

Dirk Ostwald, Evgeniya Kirilina, Ludger Starke, Felix Blankenburg

This document comprises the mathematical underpinnings of the general framework laid out in the tutorial's main text. It represents an attempt to discuss the variational Bayes approach to latent linear stochastic differential equations in their linear Gaussian state space model approximation in a mathematical coherent form. This project started out from a very modest aim: to review the variational Bayesian estimation and evaluation of latent stochastic dynamical systems in its simplest form – the linear case. During the realization of the project, however, it became more and more apparent, that, on the one hand, the “mathematization” of this nascent field is still in its infancy, and many concepts, theorems, and proofs are only brought under one umbrella by creating their formal links *de novo*. On the other hand, some more established aspects of the framework (such as Kalman-Rauch-Tung-Striebel smoothing) are so engrained in the literature, that for many authors their derivation appears to be a needless endeavor. In other words: over the course of its development, the project assumed a more and more “creative” rather than “collective” character.

This comes at a price: rather than being a collection of well-established and impermeable mathematical results, there are many occasions on which the authors had to rely on their personal analytical skills. As Howard Raiffa and Robert Schlaifer put it in the “Preface and Introduction” to their *Applied Statistical Decision Theory* (1961) The MIT Press (p.xvii): “The reader [of this supplementary material] may feel that there is an overabundance of proofs and he [or she] may well be right in so feeling. We do, however, have two things to say in self-defence. First, many of the formulas we give are hard or impossible to find in other books, and being only moderately sure of our own algebra and calculus we wanted to make it as easy as possible for the reader to verify our results; corrections will be greatly appreciated. Second, as regards the results which are perfectly well known, our desire to produce an introduction to [Variational] Bayesian analysis [for latent linear stochastic time-series models] which would be accessible even to reader who were not professionally trained in classical statistics [or time-series analysis] seemed to us to imply that we should at least give references of the proofs; but when we tried to to so, we quickly came to the conclusion that differences both in notation and in point of view (e.g., as regards to compound distributions [!]) would make such references virtually incomprehensible except to those readers who had no need of proofs anyway.” This says it all, and to emphasize one aspect again: corrections are greatly appreciated, for example, by means of email to the corresponding author. For simplicity, we refer to the high-level introductory text (to be published in the *Journal of Mathematical Psychology* as “the tutorial” in these notes, while formally, the tutorial proper comprises both the high-level introductory text and the Supplementary Material.

1 On Notation

In this section, conventions for the mathematical notation used in the tutorial and supplementary material are discussed. In general, standard mathematical notation is employed, with some concessions to the simplified notations commonly used in statistics and machine learning. The general aim of the notation has been the attempt to balance mathematical rigor and practical applicability of the presented framework.

1.1 Sets and Mappings

For analytical discussions sets S that comprise elements sharing a set-specific property p are denoted in the form

$$S = \{s | s \text{ has the property } p\} \quad (1.1)$$

A function (mapping) f is generally specified in the form

$$f: D \rightarrow R, x \mapsto f(x) \quad (1.2)$$

where the set D denotes the domain of the function f , the set R denotes the range of the function f , and ' $\mapsto$ ' denotes the mapping of the domain element $x \in D$ onto the range element $f(x) \in R$. In concordance with contemporary mathematical language, the terms 'function' and 'mapping' are used interchangeably in this tutorial. It is important to note that this (mathematical) notation distinguishes between the function f proper, and elements $f(x)$ of its range. Usually in statements as the above, the specification of the function abbreviation, its domain and its range, is followed by a definition of the functional form linking the domain elements to the range elements (for example for a quadratic function: $f(x) := x^2$).

1.2 Derivatives and Integrals

The derivative of a function of a single variable is usually denoted in 'operator' form as

$$\frac{d}{dx} f: D \rightarrow R, x \mapsto \frac{d}{dx} f(x) \quad (1.3)$$

where

$$\frac{d}{dx} f(x)|_{x=a} \quad (1.4)$$

denotes the value of the derivative in $x = a$. The partial derivative of a function of multiple variables x_i , $i = 1, \dots, n$ with respect to variable x_i is denoted as

$$\frac{\partial}{\partial x_i} f: D \rightarrow R, x \mapsto \frac{\partial}{\partial x_i} f(x) \quad (1.5)$$

Sometimes the abbreviated bar notation for the derivative, $f'(x)$, is also used.

Integrals of a function f are usually denoted as

$$\int_a^b f(\xi) d\xi \quad (1.6)$$

Often, if the domain of integration is left unspecified, or implicitly understood as comprising the entire real line, the integral boundaries a, b are omitted. Likewise, if the variable of integration is n -dimensional, the respective n -time integration is left implicit, e.g. for a function $f: \mathbb{R}^3 \rightarrow \mathbb{R}$, the integral on $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3] \subset \mathbb{R}^3$

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} f(x_1, x_2, x_3) dx_1 dx_2 dx_3 \quad (1.7)$$

is usually denoted as

$$\int f(x)dx \quad (1.8)$$

1.3 Probabilistic concepts

With respect to probabilistic concepts, the applied probabilistic notation prevalent in physics, machine learning, and statistics is employed. Specifically, probability spaces comprising an outcome set Ω , a σ -algebra $\mathcal{A}$ and a probability measure P , are not explicitly defined. Likewise, random variables will usually not be defined as measurable mappings from one measure space to another in the form

$$X: (\Omega, \mathcal{A}) \rightarrow (\Omega', \mathcal{A}'), \omega \mapsto X(\omega) \quad (1.9)$$

Rather, the exposition focuses on image measures, or the distributions of random variables, written as $P(X) := P_X(X)$, and generally defined by

$$P(X) := P_X\{X \in A'\} := P\{\omega \in \Omega | X(\omega) \in A'\} \quad (A \in \mathcal{A}') \quad (1.10)$$

For example, the statement

$$p(x = 1) = 0.1 \quad (1.11)$$

should be read as ‘the probability that the random variable x takes on a value of 1 is 0.1’. More generally, arbitrary values that a random value may take on are denoted by a superscripted asterisk, e.g. $p(x = x^*)$. An explicit distinction between probabilistic statements as frequency limits or degrees of beliefs is omitted. In concordance with the notations in machine learning textbooks, the explicit differentiation between probability distributions, probability density functions, and probability mass functions is omitted and non-capital letters are used to denote random variables and probabilities. Given the focus on Gaussian random variables in this tutorial, all probability distributions employed will actually be probability density functions, and the learning of the parameters of probability distributions should be understood as probability density function parameter learning. Finally, while some univariate examples are discussed, the theory is in general developed for random vectors, i.e. vectors of random variables of the form $x = (x_1, \dots, x_d)^T$ where the entries $x_1, \dots, x_d$ represent random variables, and $d \in \mathbb{N}$ the dimensionality of the vector. This treatment subsumes the random variable case with $d = 1$. The most important aspect of this is to keep in mind that the covariances of vectors are not positive scalars as in the random variable case, but symmetric positive-semidefinite matrices. Occasionally, colloquial reference will be made to “random variables” in cases where the random vector concept is equally appropriate.

As a notational example, the following conventions for writing bivariate and marginal probability distributions of random vectors x and y are used:

$$p(x, y): \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+, (x = x^*, y = y^*) \mapsto p(x = x^*, y = y^*) \quad (1.12)$$

for a bivariate probability distribution, where $\mathcal{X}$ and $\mathcal{Y}$ denote the ranges of x and y , respectively, and

$$p(x): \mathcal{X} \rightarrow \mathbb{R}_+, x = x^* \mapsto p(x = x^*) = \int p(x = x^*, y) dy \quad (1.13)$$

for the corresponding marginal distribution over x . With respect to conditional probability distributions, denoted here as,

$$p(x|y): \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+, (x = x^*, y = y^*) \mapsto p(x = x^* | y = y^*) = \frac{p(x=x^*, y=y^*)}{\int p(x, y=y^*) dx} \quad (1.14)$$

an important notational convention is employed to distinguish between probability distributions parameterized by fixed quantities, i.e. ‘nonrandom’ variables and probability distributions conditioned on other random variables.

Specifically, $p_\theta(x)$ is written for a probability distribution over the random variable x , which is parameterized by the fixed (nonrandom) variable θ . Likewise, $p_\vartheta(x|\theta)$ is written for a probability distribution that is conditioned on the random variable θ and parameterized by the fixed (nonrandom) variable ϑ . Finally, in the case of multivariate probability distributions, conditional distributions over subsets of variables are often referred to as conditional marginal distributions.

Due to the fact that the parameters of a Gaussian distribution are equivalent with its first central moments, namely its expectation and covariance, frequent use is made of expectation and covariance operations on random vectors. Here, the notations $\mathbb{E}_{p(x)}(f(x))$ and $\langle f(x) \rangle_{p(x)}$ are used interchangeably to denote the expectation operation of a function f of the random vector $x \in \mathcal{X}$ under the probability distribution p :

$$\langle \cdot \rangle: \mathcal{F}(x) \times \mathcal{P}(x) \rightarrow \mathbb{R}, (p(x), f(x)) \mapsto \langle f(x) \rangle_{p(x)} := \int f(x)p(x)dx \quad (1.15)$$

and

$$\mathbb{E}(\cdot): \mathcal{F}(x) \times \mathcal{P}(x) \rightarrow \mathbb{R}, (p(x), f(x)) \mapsto \mathbb{E}_{p(x)}(f(x)) := \int f(x)p(x)dx \quad (1.16)$$

where $\mathcal{F}(x)$ and $\mathcal{P}(x)$ denote unspecified¹ function and probability distribution sets over x and the integration is performed over the range $\mathcal{X}$ of x . The reason for the use of these interchangeable notations is that (in the opinion of the authors) the bracketed notation $\langle f(x) \rangle_{p(x)}$ emphasizes the ‘operator’ characteristic of an expectation slightly more, and is hence used mainly in algebraic manipulations, while the notation $\mathbb{E}_{p(x)}(f(x))$ stresses the ‘single value’ characteristic of an expectation or conditional expectation slightly more and is hence used predominantly in the statement of results rather than calculations. Covariance operations are usually denoted by the symbol $\mathbb{C}$ in the form

$$\mathbb{C}(\cdot, \cdot): \mathcal{P}(x, y) \rightarrow \mathbb{R}_+, p(x, y) \mapsto \mathbb{C}_{p(x, y)}(x, y) := \mathbb{E}_{p(x, y)} \left(\left(x - \mathbb{E}_{p(x)}(x) \right) \left(y - \mathbb{E}_{p(y)}(y) \right)^T \right) \quad (1.17)$$

which for a random vector of dimensionality $d \in \mathbb{N}$ refers to a symmetric, positive-semidefinite matrix of dimensionality $d \times d$. The probability distribution subscripts in the abbreviations $\langle x \rangle_{p(x)}$, $\mathbb{E}_{p(x)}(x)$ and $\mathbb{C}_{p(x, y)}(x, y)$ are also often omitted, if the respective distributions is clear from the context.

For random vectors x, y , i.e. vectors of random variables $x = (x_1, \dots, x_d)^T$ and $y = (y_1, \dots, y_d)^T$, $d \in \mathbb{N}$ and scalars $a, b, c \in \mathbb{R}$, some standard rules for manipulations involving expectations, variances and covariances are listed below.

Expectation of an expectation of a RV

$$\mathbb{E}(\mathbb{E}(x)) = \mathbb{E}(x) \quad (1.18)$$

Expectation of the product of a scalar and a RV

$$\mathbb{E}(ax) = a\mathbb{E}(x) \quad (1.19)$$

Expectation of the sum of two RVs

$$\mathbb{E}(x + y) = \mathbb{E}(x) + \mathbb{E}(y) \quad (1.20)$$

¹ The spaces $\mathcal{F}$ and $\mathcal{P}$ are left unspecified to eschew a formal discussion of measure theoretic concepts. The same reasoning is behind the simplified probability concept notations discussed.

Alternative expression for the variance of a RV

$$\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \quad (1.21)$$

Alternative expression for the covariance of RVs x, y

$$\mathbb{C}(x, y) = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y) \quad (1.21)$$

Covariance of a RV x with itself

$$\mathbb{C}(x, x) = \mathbb{V}(x) \quad (1.22)$$

Variance of a RV under scalar multiplication and addition

$$\mathbb{V}(ax + c) = a^2\mathbb{V}(x) \quad (1.23)$$

Note that (1.18) – (1.23) are readily transferred to the random vector case, if the corresponding rules of vector calculus are respected.

Finally, in the notation of Gaussian distributions the letter μ denotes the expectation parameter, σ and Σ denote the (co)variance parameter, and λ and Λ denote precision (inverse variance) parameters. Here the convention is followed that matrices are represented by capital letters and scalars by lower case letters. The Gaussian distribution probability density function themselves is denoted by $N(x; \mu, \Sigma)$, where x indicates the Gaussian random vector, and μ, Σ its parameters (see Supplement Section 2 for details).

1.4 Statistical concepts

With respect to the notation of statistical concepts, we note that we will mainly be dealing with iterative estimation schemes. We will denote the value that an estimated parameter takes on the i th iteration of some iterative parameter estimation scheme is denoted by a superscript (i) , such as $\theta^{(i)}, \mu^{(i)}, A^{(i)}, \Sigma_x^{(i)}$. The superscript notation thus makes the usual ‘hat’ notation for parameter estimates ($\hat{\theta}$, for example) redundant, which is hence omitted.

Table 1 summarizes the notational conventions used in the tutorial and supplementary material.

Notation	Meaning
$f: D \rightarrow R, x \mapsto f(x)$	f denotes a mapping of elements x in its domain D onto elements $f(x)$ of its range R .
$\frac{d}{dx}f: D \rightarrow R, x \mapsto \frac{d}{dx}f(x)$	$\frac{d}{dx}f$ denotes the derivative of f , which corresponds to a mapping of x onto $\frac{d}{dx}f(x)$.
$\frac{\partial}{\partial x_i}f: D \rightarrow R, x \mapsto \frac{\partial}{\partial x_i}f(x)$	For functions of multiple variables $x_1, \dots, x_n$ (e.g., $D \subset \mathbb{R}^n$), $\frac{\partial}{\partial x_i}f$ denotes the partial derivative of f with respect to x_i .
$\int_a^b f(\xi) d\xi$	Integral notation a, b indicate the lower and upper boundaries and ξ the integration variable.
$p(x, y): \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ $(x = x^*, y = y^*) \mapsto p(x = x^*, y = y^*)$	Joint probability distribution notation as prevalent in applied mathematical texts. In the context of this tutorial, $p(x, y)$ usually refers to a probability density function.
$p(x): \mathcal{X} \rightarrow \mathbb{R}_+$ $x = x^* \mapsto p(x = x^*) := \int p(x = x^*, y) dy$	Marginal probability distribution as obtained by integration ("marginalization") of continuous probability distributions.
$p(x y): \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ $(x = x^*, y = y^*) \mapsto p(x = x^* y = y^*) := \frac{p(x = x^*, y = y^*)}{\int p(x, y = y^*) dx}$	Definition of conditional probability distributions. The conditional distribution of x given $y = y^*$ equals the normalized joint probability distribution of x and $y = y^*$.
$\langle \cdot \rangle: \mathcal{F}(x) \times \mathcal{P}(x) \rightarrow \mathbb{R}$ $(p(x), f(x)) \mapsto \langle f(x) \rangle_{p(x)} := \int f(x) p(x) dx$	Notation of the expectation operation for functions f of random variables x .
$\mathbb{E}(\cdot): \mathcal{F}(x) \times \mathcal{P}(x) \rightarrow \mathbb{R}$ $(p(x), f(x)) \mapsto \mathbb{E}_{p(x)}(f(x)) := \int f(x) p(x) dx$	Alternative notation of the expectation for functions f of random variables.
$\mathbb{C}(\cdot, \cdot): \mathcal{P}(x, y) \rightarrow \mathbb{R}_+$ $p(x, y) \mapsto \mathbb{C}_{p(x, y)}(x, y) := \mathbb{E}_{p(x, y)}\left(\left(x - \mathbb{E}_{p(x)}(x)\right)\left(y - \mathbb{E}_{p(y)}(y)\right)^T\right)$	Notation of the covariance for random variables.
$\mu, \sigma, \Sigma, \lambda,$	Symbols for the parameters of Gaussian distributions.
$\mu_{x y}, \Sigma_{x y}, \sigma_{x y}$	Notation for "conditional parameters." The parameter of interest is subscripted with the random variable whose distribution it governs (x) and which is conditioned on another random variable (y).
$\theta^{(i)}, \mu^{(i)}, A^{(i)}, \Sigma_x^{(i)}$	Symbols for the value of parameter estimates on the i th iteration of an iterative parameter estimation algorithm.
$\Sigma \succ 0$	Σ is a positive-definite matrix
$\mathcal{C}^k(\mathbb{R})$	The space of k -times continuously differentiable functions on $\mathbb{R}$
$ \Sigma $	Determinant of the matrix Σ
$x \sim p(x)$	x is distributed according to $p(x)$

Table 1 Notational conventions of the tutorial and supplement

2 Some Properties of Gaussian and Gamma Distributions

The analytical properties of multivariate Gaussian distributions are of central importance for the VB approach for LGSSMs. In this section, we summarize some of these properties in theorem form. Additionally, an overview is provided in Table 2 and Figure 1.

2.1 Definition and central moments of the Gaussian

A Gaussian distribution over a random vector, i.e., a vector of random variables, $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ is defined by the probability density function

$$p_{\mu, \Sigma}(x) := N(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \quad (2.1)$$

Here, $\mu \in \mathbb{R}^d$ and the positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, $\Sigma \succ 0$ are the fixed parameters of the distribution as indicated by the subscript and semicolon notation in $p_{\mu, \Sigma}(x)$ and $N(x; \mu, \Sigma)$, respectively. If μ and Σ are themselves random variables, the statement above is written $p(x|\mu, \Sigma) = N(x|\mu, \Sigma)$ and takes on its parameterized form only for concrete values of the random variables $\mu = \mu^*$ and $\Sigma = \Sigma^*$, i.e. $p(x|\mu = \mu^*, \Sigma = \Sigma^*) = N(x; \mu^*, \Sigma^*)$. For the case of a univariate random variable $x \in \mathbb{R}$, the probability density function is usually expressed as

$$p_{\mu, \sigma^2}(x) := N(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \quad (2.2)$$

with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$. The first central moment or the expectation of a normal distribution is given by

$$\mathbb{E}_{p_{\mu, \Sigma}(x)}(x) = \langle x \rangle_{p_{\mu, \Sigma}(x)} = \mu \quad (2.3)$$

i.e., the parameter $\mu \in \mathbb{R}^d$ of a multivariate normal distribution $N(x; \mu, \Sigma)$ is equivalent to the expectation of the identity function of x under the probability distribution $N(x; \mu, \Sigma)$. The second central moment or the covariance of a normal distribution is given by

$$\mathbb{C}_{p_{\mu, \Sigma}(x)}(x, x) = \mathbb{E}_{p_{\mu, \Sigma}(x)}\left(\left(x - \mathbb{E}_{p_{\mu, \Sigma}(x)}(x)\right)\left(x - \mathbb{E}_{p_{\mu, \Sigma}(x)}(x)\right)^T\right) = \Sigma \quad (2.4)$$

i.e., the parameter $\Sigma \in \mathbb{R}^{d \times d}$, $\Sigma \succ 0$ of a multivariate normal distribution $N(x; \mu, \Sigma)$ is equivalent to the covariance of the identity function of x under the probability distribution $N(x; \mu, \Sigma)$. These two equivalences allow the expectation of the square of x under $p_{\mu, \Sigma}(x)$ to be expressed in terms of the expectation and covariance of x under $p_{\mu, \Sigma}(x)$, an important property, which is exploited on several occasions in later sections. Specifically, because of

$$\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \quad (2.5)$$

one has for normally distributed random vectors $x \in \mathbb{R}^d$:

$$\begin{aligned} \mathbb{E}_{p_{\mu, \Sigma}(x)}(xx^T) &= \mathbb{V}_{p_{\mu, \Sigma}(x)}(x) + \mathbb{E}_{p_{\mu, \Sigma}(x)}(x)\mathbb{E}_{p_{\mu, \Sigma}(x)}(x)^T \\ &= \mathbb{C}_{p_{\mu, \Sigma}(x)}(x, x) + \mathbb{E}_{p_{\mu, \Sigma}(x)}(x)\mathbb{E}_{p_{\mu, \Sigma}(x)}(x)^T \\ &= \left\langle \left(x - \langle x \rangle_{p_{\mu, \Sigma}(x)}\right) \left(x - \langle x \rangle_{p_{\mu, \Sigma}(x)}\right)^T \right\rangle_{p_{\mu, \Sigma}(x)} + \langle x \rangle_{p_{\mu, \Sigma}(x)} \langle x \rangle_{p_{\mu, \Sigma}(x)}^T \\ &= \Sigma + \mu\mu^T \end{aligned} \quad (2.6)$$

2.2 The precision parameterization

For practical purposes univariate Gaussian distributions are often parameterized using a precision parameter instead of a variance parameter. The precision is defined as the inverse of the variance $\lambda := 1/\sigma^2$. This yields the equivalent notation:

$$p_{\mu,\lambda} = N(x; \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2}(x-\mu)^2} \quad (2.7)$$

2.3 The completing-the-square theorem for the Gaussian distribution

The completing-the-square theorem states that if $x \in \mathbb{R}^d, b \in \mathbb{R}^d, A \in \mathbb{R}^{d \times d}, A \succ 0$, then

$$\exp\left(-\frac{1}{2}x^T A x + b^T x\right) = N(x; A^{-1}b, A^{-1}) \sqrt{|(2\pi A)^{-1}|} \exp\left(\frac{1}{2}b^T A^{-1}b\right) \quad (2.8)$$

The completing-the-square theorem allows identifying the parameters of a Gaussian probability density function of $x \in \mathbb{R}^d$ based on a quadratic form $x \in \mathbb{R}^d$, given by the argument of the exponent on the left hand side of the theorem.

○ It is verified by substitution of $N(x; A^{-1}b, A^{-1})$ on the right-hand side of the identity above as follows:

$$\begin{aligned} & N(x; A^{-1}b, A^{-1}) \sqrt{|(2\pi A)^{-1}|} \exp\left(\frac{1}{2}b^T A^{-1}b\right) \\ &= \frac{\sqrt{|(2\pi A)^{-1}|}}{\sqrt{|(2\pi A)^{-1}|}} \exp\left(-\frac{1}{2}(x - A^{-1}b)^T A (x - A^{-1}b) + \frac{1}{2}b^T A^{-1}b\right) \\ &= \exp\left(-\frac{1}{2}\left(x^T A x - x^T A A^{-1}b - b^T A^{-1T} A x + b^T A^{-1T} A A^{-1}b\right) + \frac{1}{2}b^T A^{-1}b\right) \\ &= \exp\left(-\frac{1}{2}\left(x^T A x - x^T b - b^T A^{-1} A x + b^T A^{-1}b\right) + \frac{1}{2}b^T A^{-1}b\right) \\ &= \exp\left(-\frac{1}{2}x^T A x + \frac{1}{2}b^T x + \frac{1}{2}b^T x - \frac{1}{2}b^T A^{-1}b + \frac{1}{2}b^T A^{-1}b\right) \\ &= \exp\left(-\frac{1}{2}x^T A x + b^T x\right) \end{aligned} \quad (2.9)$$

□

2.4 The linear transformation theorem for the Gaussian distribution

The “linear transformation theorem” states that if $x \sim N(x; \mu_x, \Sigma_x)$, where $x, \mu_x \in \mathbb{R}^d, \Sigma_x \in \mathbb{R}^{d \times d}, \Sigma_x \succ 0$, $\varepsilon \sim N(\varepsilon; \mu_\varepsilon, \Sigma_\varepsilon)$, where $\varepsilon, \mu_\varepsilon \in \mathbb{R}^d, \Sigma_\varepsilon \succ 0$, $\mathbb{C}(x, \varepsilon) = (\mathbb{C}(\varepsilon, x))^T = 0 \in \mathbb{R}^{d \times d}$ and $A \in \mathbb{R}^{d \times d}$ is a matrix, then

$$Ax + \varepsilon =: y \sim N(y; \mu_y, \Sigma_y) \quad (2.10)$$

where $y \in \mathbb{R}^d, \mu_y \in \mathbb{R}^d, \Sigma_y \in \mathbb{R}^{d \times d}, \Sigma_y \succ 0$ and specifically,

$$\mu_y = A\mu_x + \mu_\varepsilon \text{ and } \Sigma_y = A\Sigma_x A^T + \Sigma_\varepsilon \quad (2.11)$$

The proof that y is indeed a Gaussian random vector capitalizes on the transformation theorem for probability density functions and is suppressed here for brevity.

○ Verification of equations (2.10) and (2.11)

If it is acknowledged that y is a Gaussian random vector, the expressions for the parameters of this Gaussian can be derived in an instructive manner from the fact that the parameters of a Gaussian correspond to its first two central moments and the general rules for manipulating expectancies and covariances as follows:

$$\mu_y := \mathbb{E}(y) = \mathbb{E}(Ax + \varepsilon) = \mathbb{E}(Ax) + \mathbb{E}(\varepsilon) = A\mathbb{E}(x) + \mathbb{E}(\varepsilon) = A\mu_x + \mu_\varepsilon \quad (2.12)$$

and

$$\begin{aligned} \Sigma_y &:= \mathbb{E}\left((y - \mathbb{E}(y))(y - \mathbb{E}(y))^T\right) \\ &= \mathbb{E}\left((Ax + \varepsilon - \mathbb{E}(Ax + \varepsilon))(Ax + \varepsilon - \mathbb{E}(Ax + \varepsilon))^T\right) \\ &= \mathbb{E}\left((Ax + \varepsilon - A\mathbb{E}(x) - \mathbb{E}(\varepsilon))(Ax + \varepsilon - A\mathbb{E}(x) - \mathbb{E}(\varepsilon))^T\right) \\ &= \mathbb{E}\left((A(x - \mathbb{E}(x)) + (\varepsilon - \mathbb{E}(\varepsilon)))(A(x - \mathbb{E}(x)) + (\varepsilon - \mathbb{E}(\varepsilon)))^T\right) \\ &= \mathbb{E}\left((A(x - \mathbb{E}(x)) + (\varepsilon - \mathbb{E}(\varepsilon)))((x - \mathbb{E}(x))^T A^T + (\varepsilon - \mathbb{E}(\varepsilon))^T)\right) \\ &= A\mathbb{E}\left((x - \mathbb{E}(x))(x - \mathbb{E}(x))^T\right)A^T + A\mathbb{E}\left((x - \mathbb{E}(x))(\varepsilon - \mathbb{E}(\varepsilon))^T\right) \\ &\quad + \mathbb{E}\left((\varepsilon - \mathbb{E}(\varepsilon))(x - \mathbb{E}(x))^T\right)A^T + \mathbb{E}\left((\varepsilon - \mathbb{E}(\varepsilon))(\varepsilon - \mathbb{E}(\varepsilon))^T\right) \\ &= A\mathbb{C}(x, x)A^T + A\mathbb{C}(x, \varepsilon) + \mathbb{C}(\varepsilon, x)A^T + \mathbb{C}(\varepsilon, \varepsilon) \\ &= A\Sigma_x A^T + A \cdot 0 + 0 \cdot A^T + \Sigma_\varepsilon \\ &= A\Sigma_x A^T + \Sigma_\varepsilon \end{aligned} \quad (2.13)$$

where in the penultimate line the independence property $\mathbb{C}(x, \varepsilon) := 0$ of x and ε was exploited.

□

2.5 The marginalization and conditioning theorem for the Gaussian distribution

The marginalization and conditioning theorem states that, if $z \sim N(z; \mu, \Sigma)$, $z, \mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$, $\Sigma \succ 0$ and

$$z := \begin{pmatrix} x \\ y \end{pmatrix}, x \in \mathbb{R}^m, y \in \mathbb{R}^{d-m} \quad (2.14)$$

with corresponding partitions of the parameters and precision matrix $\Lambda = \Sigma^{-1} \in \mathbb{R}^{d \times d}$, i.e., with

$$\mu := \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix} \quad (2.15)$$

then

$$p(x) = N(x; \mu_x, \Sigma_{xx}) \quad (2.16)$$

Further,

$$p(x|y) = N(x; \mu_{x|y}, \Sigma_{x|y}) \quad (2.17)$$

where

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \text{ and } \Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \quad (2.17)$$

○ Preliminaries

To verify the marginalization and conditioning theorem, we follow (Bishop, 2007). The expressions for the conditional mean and conditional covariance can be verified using two intermediate steps. We first obtain some helpful relations between the elements of matrices Σ and Λ . Using the fact that $\Lambda = \Sigma^{-1}$ we have:

$$\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix} = I \quad (2.18)$$

Based on (2.18) one can infer the following identities

$$\Lambda_{xx} - \Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{yx} = \Sigma_{xx}^{-1} \quad (2.19)$$

$$\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} = \Lambda_{xx}^{-1} \quad (2.20)$$

and

$$\Lambda_{yy}^{-1}\Lambda_{yx} = -\Sigma_{yx}\Sigma_{xx}^{-1} \quad (2.21)$$

We next consider the joint probability distribution

$$p(x, y) = p(z) = N(z; \mu, \Sigma) \quad (2.22)$$

and define

$$a := (x - \mu_x) \text{ and } b := (y - \mu_y) \quad (2.23)$$

Then the quadratic form in the exponent of $p(x, y)$ can be rewritten as:

$$\begin{aligned} -\frac{1}{2}(z - \mu)^T \Sigma^{-1}(z - \mu) &= -\frac{1}{2} \begin{pmatrix} a^T & b^T \end{pmatrix} \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \\ &= -\frac{1}{2} (a^T \Lambda_{xx} a + a^T \Lambda_{xy} b + b^T \Lambda_{yx} a + b^T \Lambda_{yy} b) \\ &= -\frac{1}{2} (a^T \Lambda_{xx} a + a^T \Lambda_{xy} \Lambda_{yy}^{-1} \Lambda_{yy} b + b^T \Lambda_{yy} \Lambda_{yy}^{-1} \Lambda_{yx} a + b^T \Lambda_{yy} b) \\ &= -\frac{1}{2} a^T (\Lambda_{xx} - \Lambda_{xy} \Lambda_{yy}^{-1} \Lambda_{yx}) a - \frac{1}{2} (b + \Lambda_{yy}^{-1} \Lambda_{yx} a)^T \Lambda_{yy} (b + \Lambda_{yy}^{-1} \Lambda_{yx} a) \end{aligned} \quad (2.24)$$

Further, using the relations (2.18) – (2.21), (2.24) can be rewritten as either

$$-\frac{1}{2}(z - \mu)^T \Sigma^{-1}(z - \mu) = -\frac{1}{2}(x - \mu_x)^T \Sigma_{xx}^{-1}(x - \mu_x) - \frac{1}{2} \left(y - \mu_y - \Sigma_{yx} \Sigma_{xx}^{-1}(x - \mu_x) \right)^T \Lambda_{yy} \left(y - \mu_y - \Sigma_{yx} \Sigma_{xx}^{-1}(x - \mu_x) \right) \quad (2.25)$$

or

$$-\frac{1}{2}(z - \mu)^T \Sigma^{-1}(z - \mu) = -\frac{1}{2}\left(x - \mu_x - \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)\right)^T \Lambda_{xx}\left(x - \mu_x - \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)\right) - \frac{1}{2}(y - \mu_y)^T \Sigma_{yy}^{-1}(y - \mu_y) \quad (2.26)$$

□

○ Verification of (2.15)

To obtain the marginal distribution $p(x)$, the joint distribution $p(x, y)$ has to be integrated over y . To this end, we make use of (2.25)

$$\begin{aligned} p(x) &= \int p(x, y) dy \quad (2.27) \\ &= (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_x)^T \Sigma_{xx}^{-1}(x - \mu_x)\right) \\ &\quad \cdot \int \exp\left(-\frac{1}{2}\left(y - \mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)\right)^T \Lambda_{yy}\left(y - \mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x)\right)\right) dy \\ &= (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x - \mu_x)^T \Sigma_{xx}^{-1}(x - \mu_x)} (2\pi)^{\frac{m}{2}} |\Lambda_{yy}|^{-\frac{1}{2}} \\ &= N(x; \mu_x, \Sigma_{xx}) \end{aligned}$$

This provides verification of the statement (2.15).

□

○ Verification of (2.16)

We now consider the joint probability distribution $p(x, y = y^*)$, which, upon normalization, corresponds to the conditional distribution $p(x|y = y^*)$, because, by definition

$$p(x|y = y^*) := \frac{p(x, y = y^*)}{p(y = y^*)} \quad (2.28)$$

Rewriting the conditional distribution using quadratic form (2.26) and summarizing all terms independent of x in a constant C , one obtains

$$p(x|y = y^*) = C \cdot \exp\left(-\frac{1}{2}\left(x - \mu_x - \Sigma_{xy}\Sigma_{yy}^{-1}(y^* - \mu_y)\right)^T \Lambda_{xx}\left(x - \mu_x - \Sigma_{xy}\Sigma_{yy}^{-1}(y^* - \mu_y)\right)\right) \quad (2.29)$$

Assuming appropriate normalization we see that conditional distribution has a Gaussian form with parameters

$$\Sigma_{x|y^*} := \Lambda_{xx}^{-1} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \quad (2.30)$$

and

$$\mu_{x|y^*} := \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y^* - \mu_y) \quad (2.31)$$

□

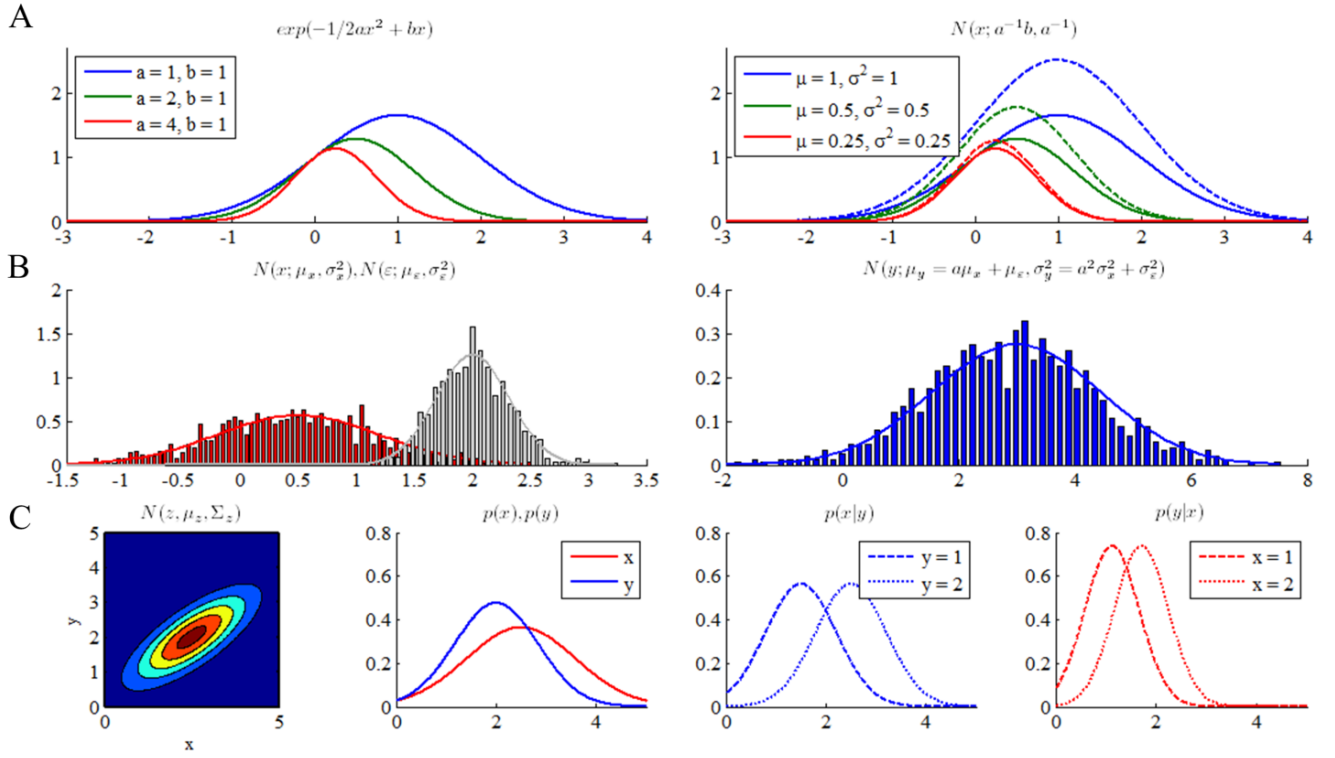


Figure 1. Visualization of properties of the Gaussian distribution. Figure 1A visualizes the completing-the-square theorem. The left-hand panel depicts three different exponential quadratic forms in x with parameters a and b as noted in the legend. For the same parameter settings, the right-hand panel depicts the evaluation of the first multiplicative term of the right-hand side of the completing-the-square theorem, i.e., the Gaussian form with parameters $\mu = a^{-1}b$ and $\sigma^2 = a^{-1}$ as dashed curves. The continuous curves represent these forms upon multiplication with the normalization factor, rendering the ensuing curves equivalent to the exponential quadratic forms in x depicted in the left-hand panel. Figure 1B visualizes the linear transformation theorem both on the analytical and realized level. The left-hand panel depicts the functional forms of two Gaussian distributions over variables x and ε as red and grey continuous curves, respectively. Additionally, a histogram estimate of both density functions based on $N = 1000$ samples is depicted as a bar graph. The right-hand panel depicts the ensuing functional form of the distribution of the random variable $x + \varepsilon$ as well as the result of a histogram estimate of the addition of the samples obtained for the left-hand panel. Figure 1C visualizes the Gaussian marginalization and conditioning theorem for a bivariate Gaussian distribution over $z = (x, y)^T$. The leftmost panel depicts the functional form of the joint distribution over x and y . The marginal distributions resulting from the application of the marginalization equations are depicted in the second panel from the left. For values $y = 1$ and $y = 2$, the third panel from the left depicts the conditional marginal distributions $p(x|y)$ as dashed and dotted lines, respectively; likewise, for values $x = 1$ and $x = 2$, the rightmost panel depicts the conditional marginal distributions $p(y|x)$ as dashed and dotted lines, respectively.

Definition and notation of the multivariate Gaussian distribution for $x \in \mathbb{R}^d$

$$p_{\mu, \Sigma}(x) := N(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

Definition and notation of the univariate Gaussian distribution $x \in \mathbb{R}$

$$p_{\mu, \sigma^2}(x) := N(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

Equivalent precision parameterization of the univariate Gaussian distribution $x \in \mathbb{R}$

$$p_{\mu, \lambda} = N(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2}(x - \mu)^2}$$

Expectation and covariance equalities and notation for multivariate Gaussian distributions

$$\begin{aligned} \mathbb{E}_{p_{\mu, \Sigma}(x)}(x) &= \langle x \rangle_{p_{\mu, \Sigma}(x)} = \mu \\ \mathbb{C}_{p_{\mu, \Sigma}(x)}(x, x) &= \mathbb{E}_{p_{\mu, \Sigma}(x)}\left(\left(x - \mathbb{E}_{p_{\mu, \Sigma}(x)}(x)\right)\left(x - \mathbb{E}_{p_{\mu, \Sigma}(x)}(x)\right)^T\right) = \Sigma \end{aligned}$$

General variance property and application to the multivariate Gaussian distribution

$$\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \Rightarrow \mathbb{E}_{p_{\mu, \Sigma}(x)}(xx^T) = \mathbb{C}_{p_{\mu, \Sigma}(x)}(x, x) + \mathbb{E}_{p_{\mu, \Sigma}(x)}(x)\mathbb{E}_{p_{\mu, \Sigma}(x)}(x)^T = \Sigma + \mu\mu^T$$

“Completing-The-Square Theorem”

The completing-the-square theorem states that if $x \in \mathbb{R}^d, b \in \mathbb{R}^d, A \in \mathbb{R}^{d \times d}, A > 0$ then

$$\exp\left(-\frac{1}{2}x^T A x - b^T x\right) = N(x; A^{-1}b, A^{-1}) \sqrt{|(2\pi A)^{-1}|} \exp\left(\frac{1}{2}b^T A^{-1}b\right)$$

“Linear Transformation Theorem”

The “linear transformation theorem” states that if $x \sim N(x; \mu_x, \Sigma_x)$, where $\mu_x \in \mathbb{R}^d, \Sigma_x \in \mathbb{R}^{d \times d}, \Sigma_x > 0$ $\varepsilon \sim N(\varepsilon; \mu_\varepsilon, \Sigma_\varepsilon)$,

where $\varepsilon, \mu_\varepsilon \in \mathbb{R}^d, \Sigma_\varepsilon > 0, \mathbb{C}(x, \varepsilon) = (\mathbb{C}(\varepsilon, x))^T = 0 \in \mathbb{R}^{d \times d}$ and $A \in \mathbb{R}^{d \times d}$ is a matrix then

$$Ax + \varepsilon =: y \sim N(y; \mu_y, \Sigma_y)$$

where $y \in \mathbb{R}^d, \mu_y \in \mathbb{R}^d, \Sigma_y \in \mathbb{R}^{d \times d}, \Sigma_y$ and specifically

$$\mu_y = A\mu_x + \mu_\varepsilon \text{ and } \Sigma_y = A\Sigma_x A^T + \Sigma_\varepsilon$$

“Marginalization and Conditioning Theorem”

The marginalization and conditioning theorem states that if $z \sim N(z; \mu, \Sigma), z, \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}, \Sigma > 0$ and

$$z := \begin{pmatrix} x \\ y \end{pmatrix}, x \in \mathbb{R}^m, y \in \mathbb{R}^{d-m}$$

with corresponding partitions of the parameters and variance matrix $\Sigma \in \mathbb{R}^{d \times d}$, i.e., with

$$\mu := \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

then

$$p(x) = N(x; \mu_x, \Sigma_{xx}) \text{ and } p(y) = N(y; \mu_y, \Sigma_{yy})$$

Further,

$$p(x|y) = N(x; \mu_{x|y}, \Sigma_{x|y}), \text{ where } \mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \text{ and } \Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$$

Table 2. Some Properties of the Gaussian distribution.

2.6 Information theoretic quantities for Gaussian probability density functions

The differential entropy of random variable x governed by a probability density function $p(x)$ is defined as (Cover and Thomas (2006), p. 243)

$$\mathcal{H}(p(x)) := - \int p(x) \ln(p(x)) dx \quad (2.32)$$

Without proof, we note that for a d -dimensional random vector $x \in \mathbb{R}^d$ governed by the d -dimensional Gaussian distribution $N(x; \mu, \Sigma)$, this integral evaluates to (Cover and Thomas (2006), p. 250)

$$\mathcal{H}(N(x; \mu, \Sigma)) = \frac{1}{2} \ln |\Sigma| + \frac{d}{2} (1 + \ln 2\pi) = \frac{1}{2} \ln((2\pi e)^d |\Sigma|) \quad (2.33)$$

For the special univariate case $\in \mathbb{R}$, the above simplifies to (Cover and Thomas (2006), p. 244)

$$\mathcal{H}(N(x; \mu, \sigma^2)) = \frac{1}{2} \ln \sigma^2 + \frac{1}{2} (1 + \ln 2\pi) = \frac{1}{2} \ln(2\pi e \sigma^2) \quad (2.34)$$

As noted in Appendix A of the tutorial, the Kullback-Leibler divergence from probability density function $q(x)$ over a random variable x to a probability density function $p(x)$ over the same random variable is defined as

$$\mathcal{KL}(q(x)||p(x)) := \int q(x) \ln \left(\frac{q(x)}{p(x)} \right) dx \quad (2.35)$$

Without proof (Penny, 2001), we note that the KL divergence between two Gaussian probability density functions on a d -dimensional random vector $x \in \mathbb{R}^d$, $N(x; \mu_q, \Sigma_q)$ and $N(x; \mu_p, \Sigma_p)$, is given by

$$\mathcal{KL}(N(x; \mu_q, \Sigma_q)||N(x; \mu_p, \Sigma_p)) = \frac{1}{2} \frac{|\Sigma_p|}{|\Sigma_q|} + \frac{1}{2} \text{tr}(\Sigma_p^{-1} \Sigma_q) + \frac{1}{2} (\mu_q - \mu_p)^T \Sigma_p^{-1} (\mu_q - \mu_p) - \frac{d}{2} \quad (2.36)$$

where $\text{tr}(A) = \sum_{i=1}^d a_{ii}$ denotes the trace of a matrix $A \in \mathbb{R}^{d \times d}$. For the special univariate case $\in \mathbb{R}$, the above simplifies to

$$\mathcal{KL}(N(x; \mu_q, \sigma_q^2)||N(x; \mu_p, \sigma_p^2)) = \frac{1}{2} \frac{\sigma_p^2}{\sigma_q^2} + \frac{\mu_q^2 + \mu_p^2 - \sigma_q^2 - 2\mu_q \mu_p}{2\sigma_p^2} - \frac{1}{2} \quad (2.37)$$

2.7 The Gamma distribution

The Gamma density is considered in its “shape and scale” parameterization given as

$$G(\lambda; a, b) := \frac{1}{\Gamma(a)} \frac{1}{b^a} \lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right) \quad \text{for } \lambda > 0 \text{ and } a, b > 0 \quad (2.38)$$

where a is referred to as the “shape parameter” and b is referred to as the “scale parameter”. $\Gamma(x)$ denotes the Gamma function which is defined for $x > 0$ as:

$$\Gamma: \mathbb{R}^+ \setminus \{0\} \rightarrow \mathbb{R}, x \mapsto \Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \quad (2.39)$$

The expectation and variance of λ under $G(\lambda; a, b)$ are expressed in terms of the parameters as

$$\mathbb{E}_{G(\lambda; a, b)}(\lambda) = ab \text{ and } \mathbb{V}_{G(\lambda; a, b)}(\lambda) = ab^2 \quad (2.40)$$

The expectation of the logarithm of λ under $G(\lambda; a, b)$ is given by

$$\mathbb{E}_{G(\lambda; a, b)}(\ln \lambda) = \psi(a) + \ln b \quad (2.41)$$

where ψ denotes the digamma function. For $x > 0$ the digamma function may be presented in its integral form as

$$\psi: \mathbb{R}^+ \setminus \{0\} \rightarrow \mathbb{R}, x \mapsto \psi(x) = \int_0^\infty \left(\frac{e^{-t}}{t} - \frac{e^{-xt}}{1-e^{-t}} \right) dt \quad (2.42)$$

Without proof, we note that the differential entropy of a random variable λ governed by a Gamma probability density function in its “shape and scale” parameterization is given by

$$\mathcal{H}(G(\lambda; a, b)) = a + \ln b + \ln \Gamma(a) + (1 - a)\psi(a) \quad (2.43)$$

where Γ and ψ denote the gamma and digamma functions, respectively. Likewise, without proof (Penny, 2001), we note that the KL divergence between two Gamma probability density functions on a random variable $\lambda > 0$, $G(\lambda; a_q, b_q)$ and $G(\lambda; a_p, b_p)$, is given by

$$\begin{aligned} \mathcal{KL}(G(\lambda; a_q, b_q) \parallel G(\lambda; a_p, b_p)) \\ = (a^q - 1)\psi(a^q) - \ln b^q - a^q - \ln \Gamma(a^q) + \ln \Gamma(a^p) + a^p \ln b^p - (a^p - 1)(\psi(a^q) + \ln b^q) + \frac{a^q b^q}{b^p} \end{aligned} \quad (2.44)$$

where Γ denotes the gamma, and ψ denotes the digamma function.

3 Wiener Processes, Stochastic Integration, and Euler Maruyama Discretization

The aim of this section is to review a minimum set of concepts from stochastic analysis which forms the background for the embedding of (discrete-time) LGSSMs in the context of (continuous-time) stochastic differential equations as discussed in Sections 2.1 and 5 of the tutorial. Supplement Section 3 is not meant to provide a comprehensive discussion of stochastic analysis, but rather to collect some pointers to aspects of stochastic analysis the interested reader might pursue. Additionally, we hope to provide some interpretation for equations (1) –(6) of Section 2.1, as well as equations (36) – (37) of the tutorial example in Section 5.

3.1 Wiener processes

In this section, the intuition and mathematical formulation of Wiener processes is briefly sketched. For a comprehensive review of stochastic processes in general and Wiener processes in particular, the reader is referred to (Bass, 2011). In the context of the tutorial, Wiener processes form the mathematical model of random fluctuations and the latent level (ref. equation 1 of the tutorial), which, by means of Euler-Maruyama approximation, results in the Gaussian state space evolution of the LGSSM (ref. equation 2 of the tutorial). Here, we proceed by introducing the definition of general stochastic processes, the physical intuition of Wiener processes, the definition of Wiener processes as stochastic processes, and finally some properties of Wiener processes required for the discussion of stochastic integration.

General Stochastic Processes

Informally, a general stochastic process can be regarded as a set of random variables $(X_t)_{t \in T}$ indexed by an uncountable index set $T \subseteq \mathbb{R}_+$ that models continuous time. Formally, a general stochastic process is defined as follows: Let $(\Omega, \mathcal{A}, P)$ be a probability space and T an arbitrary nonempty set. Let $X_t: (\Omega, \mathcal{A}) \rightarrow (\mathcal{X}_t, \mathcal{B}_t)$ be a $(\mathcal{A}, \mathcal{B}_t)$ - random variable for every $t \in T$. Then

$$X_T := (\Omega, \mathcal{A}, P, (X_t)_{t \in T}) \quad (3.1)$$

denotes a general stochastic process with parameter set T .

The uncountability of the parameter set T entails some mathematical difficulties for the definition of the outcome set Ω of a stochastic process, its corresponding σ -field $\mathcal{A}$, and the existence of the probability measure P . These difficulties can be resolved using Kolmogorov's theorem (Bass, 2011), which allows for proving the existence of a stochastic process based on a consistent set of finite marginal distributions. In other words, the definition of the Wiener processes based on normally distributed increments given below can be related back to the general definition of stochastic processes by means of Kolmogorov's theorem. Below, we often use the simplified notation $X(t)$ for a general stochastic process.

Physical Intuition of Wiener Processes

It is helpful to consider the physical interpretation of a Wiener process to motivate the formal definition below. In its physical interpretation, a realization of a one-dimensional Wiener process can be regarded as the projection of the displacement vector of a three-dimensional particle in a fluid onto a selected coordinate axis of $\mathbb{R}^3$ over time. Specifically, $X_0 = 0$ would denote the particle centered on its starting point and X_t the projection of its displacement from the starting point at time t onto e.g., the x -axis. Importantly, the particle is considered large with respect to the molecules constituting the fluid. In principle, based on a set of initial conditions, the movement of all molecules in the fluid is determined by Hamiltonian (i.e., generalized Newtonian) dynamics. However, because the number of molecules is conceived as very large, a statistical or probabilistic view is appropriate. The molecules of the fluid are imagined to collide with the large particle at a time scale that is short with respect to the time scale that the particle is observed on. In this sense, the movement of the large particle in a time interval $[t_1, t_2]$ can be imagined as being the result of the summation of a large number of independent fluid-molecule-particle collisions, each evoking a minute movement of the particle. Informally, the central limit theorem of probability theorem states that the sum (or average) of a high number of independently distributed random variables (here the minute motions of the particle evoked by each molecule-particle collision) is normally distributed. Thus, displacement in X , i.e. the difference $X_{t_2} - X_{t_1}$ can be considered to be normally distributed, and, because the molecule impacts are conceived as independent and evoking motion in all spatial directions, have an expectation of zero, i.e. $E(X_{t_2} - X_{t_1}) = 0$. In other words, most likely, due to the complementary nature of the summed impacts between t_1 and t_2 , the particle stays where it is. Further, if the characteristics of the fluid do not change over time (e.g. the fluid is not heated up, causing stronger impacts of the molecules), the distribution of the spatial distance in a time interval $[t_1, t_2]$ should be the same as the distribution of the spatial distance in a time interval $[t_1 + h, t_2 + h]$, with $h > 0$. Finally, the particle motions evoked by the impact of fluid molecules are considered to be independent over disjoint time intervals.

Definition of Wiener Processes²

The intuition of a Wiener process realization as a large particle colliding with many small fluid molecules at a short time scale with respect to the observation time scale of the movement of the large particle yields the following set of formal requirements for a mathematical model $(\Omega, \mathcal{A}, P, (X_t)_{t \in T})$ of Wiener processes: $(\Omega, \mathcal{A}, P, (X_t)_{t \in T})$ is a stochastic process with parameter set $T = \mathbb{R}_+$, for which the following requirements hold:

- (1) *Initial value* $X_0 = 0$ P -almost everywhere (i.e., if there exists $N \subset \Omega$ for which $X_0(\omega_N) \neq 0, \omega_N \in N$, then $P(N) = 0$). In other words, $P\{X_0 = 0\} = P(\{\omega \in \Omega | X_0(\omega) = 0\}) = 1$.
- (2) *Normality* For every $t > 0, X_t$ is normally distributed with expectation 0 and variance parameter $v(t)$. Here $v(\cdot)$ denotes the variance of the normal distribution of X_t as a function of t .
- (3) *Stationarity and Independence* $(\Omega, \mathcal{A}, P, (X_t)_{t \in T})$ is a stochastic process with stationary increments, i.e. the distribution $P_{X_t - X_s}$ of the increments $X_t - X_s$ ($s, t \in T, s < t$) only depends on the difference $t - s$, but not on t or s itself. In other words, for $h > 0$, we have $P_{X_{t+h} - X_{s+h}} = P_{X_t - X_s}$. Note that with X_t and X_s also the increment $X_t - X_s$ is a random variable. Further, $(\Omega, \mathcal{A}, P, (X_t)_{t \in T})$ is a stochastic process with stochastically

² In the following section, $|S|$ denotes the cardinality of a set S , not a determinant.

independent increments, i.e. for all $J = \{t_1, \dots, t_n\} \in \mathcal{H}(T)$ (where $\mathcal{H}(T) := \{J | J \subset T, 0 < |J| < \infty\}$) with $n \geq 2$ and $t_1 < \dots < t_n$ the random variables (increments) $X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}}$ are stochastically independent.

Note that, formally, the mathematical framework of stochastic processes requires an existence proof of the postulated probability measure P on $(\Omega, \mathcal{A})$, which can be achieved by showing the consistency of the set of marginal distributions defined above. We eschew this existence proof here and instead note some informal properties of Wiener processes required for the concept of stochastic integration.

Properties of Wiener processes

We first reformulate the Wiener process without reference to the underlying measure space and, from now on, denote the Wiener process by $W(t), t \in [0, T]$ (Hassler, 2007). Informally, $W(t)$ is a process with a starting value of 0, and independent, normally distributed, stationary increments. More formally, we can restate (1) – (3) in the definition of the Wiener process above as follows

- (1) *Initial value* The initial value of the Wiener process is zero with probability 1, i.e. $P(W(0) = 0) = 1$.
- (2) *Normality* The increments $W(t_1) - W(t_0), \dots, W(t_n) - W(t_{n-1})$, where $0 \leq t_0 \leq t_1 \leq \dots \leq t_n$ are independent for $n \in \mathbb{N}$.
- (3) *Stationarity and Independence* The increments $W(t) - W(s)$ are normally distributed with expectation 0 and variance equal to the temporal difference $t - s$ for $0 \leq s < t$, i.e. $W(t) - W(s) \sim N(W(t) - W(s); 0, t - s)$. Notably, the variance of the increments is stationary, i.e. it only depends on the difference $t - s$, but not the absolute values of s and t , and the joint distribution of the increments is multivariate normal with a diagonal covariance matrix.

The Wiener process is defined by the properties of its increments. Nevertheless, the Wiener process can, intuitively, be conceived as a stochastic function of the form $W: [0, T] \rightarrow \mathbb{R}, t \mapsto W(t)$ where $W(t)$ is distributed according to a normal distribution $N(W(t); 0, t)$. This is readily seen by considering the distribution of the increment $W(t) - W(0)$ for $t \in [0, T]$. According to (3) $W(t) - W(0) \sim N(W(t) - W(0); 0, t - 0)$ and according to (1) $W(0) = 0$ with probability 1. Based on the properties of normal distributions, we can thus infer that $\mathbb{E}(W(t)) = 0$ and $\mathbb{V}(W(t)) = t$.

Gaussian Processes

Equivalently, a Wiener process $W(t)$ can be conceived as a Gaussian process. A Gaussian process is a stochastic process $(X_t)_{t \in T}$ with a countable index set $T := \{t_1, \dots, t_n\}$, $t_1 \leq t_2 \leq \dots \leq t_n$, where $|T| = n, n \in \mathbb{N}$, which is multivariate normally distributed. In terms of Gaussian processes, the vector $W(t) := (W(t_1), \dots, W(t_n))^T$ is distributed according to $N(W(t); 0, \Sigma)$, where the entries of the covariance matrix $\Sigma \in \mathbb{R}^{n \times n}, \Sigma$ are given by $(\Sigma)_{ij} = \min(t_i, t_j)$ (for a proof, see e.g. (Hassler, 2007)). Note in particular, that the diagonal elements of this covariance matrix are increasing and that the off-diagonal elements are non-zero, i.e. modeling stochastic dependencies.

3.2 Stochastic integration

Stochastic integrals, unlike the integrals known from standard calculus and analysis, are random variables, implying that the values they assume display some “randomness”. The aim of this section is to introduce the Ito integral of a stochastic process as the prerequisite for the framework of stochastic differential equations. To help with its introduction, the Ito integral is developed as the endpoint of a “stochastification” of the Riemann integral known from standard calculus. The development of this section is largely based on (Hassler, 2007), Sections 6 - 11.

Convergence in the quadratic mean

In this section, we lay the foundations for the definitions of stochastic integrals as the limits of their corresponding sum expressions, in analogy to the well-known quadrature expressions for Riemann integrals in deterministic analysis. To this end, we require a convergence notion for random variables and discuss the supporting interval partition used in the subsequent sections.

The convergence form required in the definition of stochastic integrals below is “convergence in the quadratic mean”. A sequence of random variables $(X_n)_{n \in \mathbb{N}}$ is said to converge to a random variable X in the quadratic mean, if $\mathbb{E}((X_n - X)^2) \rightarrow 0$ for $n \rightarrow \infty$, which will be denoted as $X_n \xrightarrow{2} X$. Without proof, we state the following (Cauchy) convergence criterion: A sequence of random variables with $\mathbb{E}(X_n^2) < \infty$ converges in the quadratic mean, if for arbitrary n and m , $\mathbb{E}((X_m - X_n)^2) \rightarrow 0$ for $m, n \rightarrow \infty$, or, equivalently, if for arbitrary n and m , $\mathbb{E}(X_m X_n) \rightarrow c < \infty$ for $m, n \rightarrow \infty$, $c \in \mathbb{R}$.

The concept of convergence in the quadratic mean can be demonstrated using a simple example: Consider the sequence $(X_n)_{n \in \mathbb{N}}$ of random variables defined by $X_n := \frac{1}{n} \sum_{i=1}^n Y_i$ for $n = 1, 2, \dots$ where the $Y_i, i = 1, \dots, n$ are independent, normally distributed random variables with mean 0 and variance σ^2 , i.e. $Y_i \sim N(Y_i; 0, \sigma^2)$ for $i = 1, \dots, n$. Using the Cauchy convergence criterion above, it can be readily shown that this sequence converges in the quadratic mean (assuming $\mathbb{E}(X_n^2) < \infty$) as follows

$$\mathbb{E}(X_m X_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n Y_i \cdot \frac{1}{m} \sum_{j=1}^m Y_j\right) = \frac{1}{mn} \sum_{i=1}^{\min(n,m)} \mathbb{E}(Y_i^2) = \sigma^2 \frac{\min(n,m)}{mn} \rightarrow 0 \quad (3.2)$$

for $n, m \rightarrow \infty$. In the above, the second equation follows, because the expectations of the products $Y_i \cdot Y_j$ for $i \neq j$ are 0, as the $Y_i, i = 1, 2, \dots$ are assumed to be stochastically independent.

Interval partitions

In order to be able to define the integral of a function on the interval $[0, T] \subset \mathbb{R}$, the interval $[0, T]$ is partitioned into disjoint subintervals of the form

$$S_n([0, T]) := [t_0, t_1] \cup [t_1, t_2] \cup \dots \cup [t_{n-1}, t_n] \quad (3.3)$$

where $0 = t_0 < t_1 < t_2 < \dots < t_{n-1} < t_n = T$. It is assumed that this partition becomes more and more fine grained as n increases, in other words: $\max_{1 \leq i \leq n} (t_i - t_{i-1}) \rightarrow 0$ for $n \rightarrow \infty$. A typical example of a partition that fulfills this criterion is the equidistant partition of $[0, T]$, given by $t_i = \frac{iT}{n}$ ($i \in \mathbb{N}_n$), for which one has

$$t_i - t_{i-1} = i \frac{T}{n} - (i-1) \frac{T}{n} = \frac{T}{n} \quad (i \in \mathbb{N}_n) \quad (3.4)$$

which obviously fulfills $\lim_{n \rightarrow \infty} \frac{T}{n} = 0$. We note that for any deterministic function $f: [0, T] \rightarrow \mathbb{R}$,

$$\begin{aligned} \sum_{i=1}^n f(t_i) - f(t_{i-1}) &= f(t_1) - f(t_0) + f(t_2) - f(t_1) + \dots + f(t_n) - f(t_{n-1}) \\ &= f(t_n) - f(t_0) \\ &= f(T) - f(0) \end{aligned} \quad (3.5)$$

and thus, for the identity function, we have

$$\sum_{i=1}^n t_i - t_{i-1} = T - 0 = T \quad (3.6)$$

An arbitrary point of support in the interval $[t_{i-1}, t_i]$ will be denoted by t_i^* .

The stochastic Riemann integral

Informally, stochastic Riemann integrals are the usual Riemann integrals with a stochastic process being part of the function integrated over. If the stochastic process in question is a Wiener process, the stochastic Riemann integral is normally distributed with an expectation of zero and a closed form equation for its variance can be obtained. Formally, the general stochastic Riemann integral over the product of a deterministic function f and a general stochastic process $X(t)$ is defined as the limit of the Riemann sum

$$R_n = \sum_{i=1}^n f(t_i^*)X(t_i^*)(t_i - t_{i-1}) \quad (3.7)$$

If the limit of this sum converges in the quadratic mean for $n \rightarrow \infty$ independently of the form of the partition $S_n([0, T])$, this limit is defined as the stochastic Riemann integral:

$$\int_0^T f(s)X(s) ds := \lim_{n \rightarrow \infty} \sum_{i=1}^n f(t_i^*)X(t_i^*)(t_i - t_{i-1}) \quad (3.8)$$

We next consider stochastic Riemann integrals, for which the stochastic process is a Wiener process. For a deterministic function f , the stochastic Riemann integral of the product between f and a Wiener process $W(t)$ corresponds to a normal distribution with expectation

$$\mathbb{E} \left(\int_0^T f(s)W(s) ds \right) = 0 \quad (3.9)$$

and covariance

$$\mathbb{C} \left(\int_0^T f(s)W(s) ds, \int_0^T f(s)W(s) ds \right) = \int_0^T \int_0^T f(s)f(t) \min(s, t) ds dt \quad (3.10)$$

(for a proof of the above, see (Hassler, 2007)).

The Riemann-Stieltjes integral

Riemann-Stieltjes integrals are solutions for specific stochastic differential equations and are normally distributed random variables. The integrand of a Riemann-Stieltjes integral is a deterministic function f , however, this is integrated with respect to a Wiener process, which defines the Riemann-Stieltjes sum as follows

$$RS_n = \sum_{i=1}^n f(t_i^*)(W(t_i) - W(t_{i-1})) \quad (3.11)$$

As for the stochastic Riemann integral, if the limit of this sum for $n \rightarrow \infty$ exists in the quadratic mean sense independently of the form of the partition $S_n([0, T])$, this limit is defined as the Riemann-Stieltjes integral

$$\int_0^T f(s) dW(s) := \lim_{n \rightarrow \infty} \sum_{i=1}^n f(t_i^*)(W(t_i) - W(t_{i-1})) \quad (3.12)$$

As a result of being the limit of a sum of normally distributed random variables the Riemann-Stieltjes integral is normally distributed with expectation

$$\mathbb{E} \left(\int_0^T f(s) dW(s) \right) = 0 \quad (3.13)$$

and covariance

$$\mathbb{C} \left(\int_0^T f(s) dW(s), \int_0^T f(s) dW(s) \right) = \int_0^T f^2(s) ds \quad (3.14)$$

(for a proof of the above, see (Hassler, 2007)). To see that for the Riemann-Stieltjes integral of the constant unity function

$$\int_{t_1}^{t_2} dW(s) = W(t_2) - W(t_1) \quad (3.15)$$

we consider its relation to the concept of partial integration as known from the calculus of standard Riemann integrals. Recall that, for two deterministic integrable functions $f: [0, T] \rightarrow \mathbb{R}$ and $g: [0, T] \rightarrow \mathbb{R}$, the following rule of integration holds

$$\int_0^T f(t)g'(t)dt = f(t)g(t)|_0^T - \int_0^T f'(t)g(t)dt \quad (3.16)$$

For the Riemann-Stieltjes integral, one has (changing the upper integral bound from T to $t \in [0, T]$, and the variable of integration from t to s , for a proof see (Hassler, 2007))

$$\int_0^t f(s)dW(s) = f(s)W(s)|_0^t - \int_0^t W(s)df(s) \quad (3.17)$$

which can be reformulated as

$$\int_0^t f(s)dW(s) = f(t)W(t) - \int_0^t W(s)f'(s)ds \quad (3.18)$$

Finally, we consider the special case of the Riemann-Stieltjes integral of the unity function $f(t) := 1$. Substitution into the equation above yields

$$\int_0^t dW(s) = W(t) - W(0) - \int_0^t W(s) \cdot 0 ds = W(t) \quad (3.19)$$

More generally, on the interval $[t_1, t_2] \subset [0, T]$, one thus has

$$\int_{t_1}^{t_2} dW(s) = W(s)|_{t_1}^{t_2} - \int_{t_1}^{t_2} W(s) \cdot 0 ds = W(t_2) - W(t_1) \quad (3.20)$$

The Ito integral

The Ito integral provides the general basis for stochastic differential equations as discussed below. Here, we briefly discuss its definition and some of its properties. The general Ito integral of a stochastic process $X(t)$ with respect to a Wiener process $W(t)$ is defined as limit of the Ito sum

$$I_n := \sum_{i=1}^n X(t_{i-1})(W(t_i) - W(t_{i-1})) \quad (3.21)$$

where, importantly (and in contrast to other stochastic integrals, as e.g., Stratonovich integral) the point t_{i-1} , i.e., the left border of the partition interval $[t_{i-1}, t_i]$ serves as supporting point. If $X(t)$ is a stochastic process with finite variance, and if $X(t)$ is independent of the future of the stochastic process, the Ito sum converges to a well-defined limit in the quadratic means sense and is defined as the Ito integral

$$\int_0^T X(t) dW(t) := \lim_{n \rightarrow \infty} \sum_{i=1}^n X(t_{i-1})(W(t_i) - W(t_{i-1})) \quad (3.22)$$

The Ito integral is a random variable and its first two central moments can be shown to be given by (for a proof see (Hassler, 2007))

$$\mathbb{E} \left(\int_0^T X(t) dW(t) \right) = 0 \quad (3.23)$$

and

$$\mathbb{C} \left(\int_0^T X(t) dW(t), \int_0^T X(t) dW(t) \right) = \int_0^T \mathbb{E}(X^2(t)) dt \quad (3.24)$$

In the degenerate case that the stochastic process $X(t)$ corresponds to the unity function, i.e., that $P(X(t) = 1) = 1$, the Ito integral evaluates to the Riemann-Stieltjes integral discussed in the previous section, which in turn

evaluates to the increment of a Wiener process (as seen above, again changing the upper integral bound from T to $t \in [0, T]$, and the variable of integration from t to s):

$$\int_0^t X(s) dW(s) = \int_0^t 1 dW(s) = W(t) \quad (3.25)$$

3.3 Stochastic Differential Equations

Diffusions and Stochastic Differential Equations (SDEs)

A diffusion is a stochastic process comprising the sum of a stochastic Riemann integral and an Ito integral in the form

$$X(t) = X(0) + \int_0^t \mu(X(s), s) ds + \int_0^t \sigma(X(s), s) dW(s) \quad (3.26)$$

where $X(0)$ indicates the starting point of the stochastic process X . Here, the functions μ and σ are referred to as a drift function and volatility functions, respectively, and are functions of time and the diffusion process itself. The first integral term, a stochastic Riemann integral, models a deterministic drift of the process X , while the second integral term, an Ito integral, models a random component. Taking the derivative with respect to time on both sides of the diffusion equation above yields its differential form

$$\frac{d}{dt} X(t) = \frac{d}{dt} \left(X(0) + \int_0^t \mu(X(s), s) ds + \int_0^t \sigma(X(s), s) dW(s) \right) \quad (3.27)$$

which is usually denoted as

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t) \quad (3.28)$$

Informally, the equation above states, that the change of the process X within an “infinitesimal” small time interval at time point t is given as the sum of (1) the product of a deterministic function μ (depending on the state of the process X at time point t and time itself) with the “infinitesimal” length of the time interval dt and (2) the product of a deterministic function σ (also depending on the state of the process X at time t and time itself) with the “infinitesimal” small increment of a Wiener process at time t , which is randomly distributed. The classification of stochastic differential equations (SDEs), as well as the study of existence and uniqueness properties of their solutions, forms an integral part of stochastic analysis. The interested reader is referred to (Kloeden & Platen, 1999) for an in depth analytical and numerical treatment of SDEs. Here, we briefly introduce the set of narrow-sense linear SDEs with the aim of introducing the Langevin equation as provided in equation (4).

Autonomous narrow-sense linear SDEs and the Langevin equation

An SDE of the type

$$dX(t) = \mu(X(t), t)dt + \sigma(X(t), t)dW(t) \quad (3.29)$$

is called a linear SDE, if the functions $\mu: \mathbb{R} \times [0, T] \rightarrow \mathbb{R}$ and $\sigma: \mathbb{R} \times [0, T] \rightarrow \mathbb{R}$ take the form

$$\mu(X(t), t) := \mu_1(t)X(t) + \mu_2(t) \text{ and } \sigma(X(t), t) := \sigma_1(t)X(t) + \sigma_2(t) \quad (3.30)$$

for coefficients $\mu_1, \mu_2, \sigma_1, \sigma_2: [0, T] \rightarrow \mathbb{R}$. If all coefficients are constants, the linear SDE is called autonomous. Further, if $\mu_2(t) = \sigma_2(t) := 0$, the linear SDE is referred to as homogenous. Finally, the linear SDE is referred to as linear in the narrow sense, if $\sigma_1(t) := 0$, or in other words, the noise term is additive. The autonomous narrow-sense linear SDE thus has the general form

$$dX(t) = (\mu_1 X(t) + \mu_2)dt + \sigma_2 dW(t) \quad (3.31)$$

Setting $\mu_1 := \alpha, \mu_2 := 0$ and $\sigma_2 := \sigma$, one obtains the “latent linear diffusion process” of equation (4) in the tutorial, commonly known as the Langevin equation:

$$dX(t) = \alpha X(t)dt + \sigma dW(t) \quad (\alpha < 0, \sigma > 0, t \in [0, s]) \quad (3.32)$$

The general solution of the Langevin equation is given by (for a proof see (Hassler, 2007))

$$X(t) = \exp(\alpha t)X(0) + \exp(\alpha t)X(0) \int_0^t \sigma \exp(s) dW(s) \quad (3.33)$$

which specifies a homogenous Gaussian process with expectation

$$\mathbb{E}(X(t)) = \exp(\alpha t) \mathbb{E}(X(0)) \quad (3.34)$$

and variance

$$\mathbb{V}(X(t)) = \exp(\alpha t) \mathbb{V}(X(0)) + \frac{\sigma^2}{2\alpha} (1 - \exp(\alpha t)) \quad (3.35)$$

3.4 Euler-Maruyama discretization

A commonly employed discretization technique for SDEs is the Euler-Maruyama method (Hassler, 2007; Kloeden & Platen, 1999), which defines a recursive scheme for the numerical evaluation of SDEs on a time interval $[0, s] \subset \mathbb{R}$. The Euler-Maruyama method is based at a set of discrete time points $t_i \in [0, s]$, $i = 0, 1, \dots, n$ with

$$0 = t_0 < t_1 = \frac{s}{n} < \dots < t_{n-1} < t_n = s$$

which represent the discretization of the observation time interval

$$[0, s] = \cup_{i=1}^n \left[\frac{i-1}{n}s, \frac{i}{n}s \right] \quad (3.36)$$

into n equisized bins of length $\Delta t := \frac{s}{n}$, and which in the context of the tutorial may be conceived as being provided by the sampling rate of the measurement or observation process. The first step towards the Euler-Maruyama approximation of (3.32) to rewrite it in its integral form given by

$$X(t) = X(0) + \int_0^t \alpha X(\tau) d\tau + \int_0^t \sigma dW(\tau) \quad t \in [0, s] \subset \mathbb{R} \quad (3.37)$$

In (3.37) the first integral term denotes a stochastic Riemann integral and the second term denotes a Riemann-Stieltjes integral. Considering this integral form on the subinterval $[t_{i-1}, t_i] \subset [0, s]$ ($i = 1, \dots, n$) then yields

$$X(t_i) = X(t_{i-1}) + \int_{t_{i-1}}^{t_i} \alpha X(\tau) d\tau + \int_{t_{i-1}}^{t_i} \sigma dW(\tau) \quad (3.38)$$

By approximating the value of $X(s)$ for $s \in [t_{i-1}, t_i]$ by its left endpoint value, that is, by setting $X(s) := X(t_{i-1})$ for all $s \in [t_{i-1}, t_i]$, one obtains the Euler-Maruyama approximation of the stochastic differential equation

$$X(t_i) = X(t_{i-1}) + \int_{t_{i-1}}^{t_i} \alpha X(t_{i-1}) d\tau + \int_{t_{i-1}}^{t_i} \sigma dW(\tau) \quad (3.39)$$

$$= X(t_{i-1}) + \alpha X(t_{i-1}) \int_{t_{i-1}}^{t_i} d\tau + \sigma \int_{t_{i-1}}^{t_i} dW(\tau) \quad (3.40)$$

for $i = 1, \dots, n$. Riemann integration of the first integral term yields

$$\int_{t_{i-1}}^{t_i} d\tau = \tau|_{t_{i-1}}^{t_i} = t_i - t_{i-1} = \frac{s}{n} = \Delta t \quad (3.41)$$

and Riemann-Stieltjes integration of the second integral yields

$$\int_{t_{i-1}}^{t_i} dW(\tau) = W(t_i) - W(t_{i-1}) \quad (3.42)$$

From the definition of Wiener processes we note that that

$$W(t_i) - W(t_{i-1}) = W\left(\frac{is}{n}\right) - W\left(\frac{(i-1)s}{n}\right) = W(\Delta t) \quad (3.45)$$

is normally distributed with expectation parameter 0 and variance parameter Δt . To rewrite

$$dX(t) = \alpha X(t)dt + \sigma dW(t) \quad (3.46)$$

in the familiar autoregressive process of order 1 (AR(1)) form

$$x_t = \alpha x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(\varepsilon, 0, \sigma_x^2) \quad (3.47)$$

we thus start by considering

$$X(t_i) = X(t_{i-1}) + \alpha X(t_{i-1})\Delta t + \sigma(W(t_i) - W(t_{i-1})) \quad (3.48)$$

and change the notation by setting

$$x_{t_i} := X(t_i), \quad w_{t_i} := W(t_i) \text{ and } \varepsilon_{t_i} := w_{t_i} - w_{t_{i-1}} \quad (3.49)$$

resulting in

$$x_{t_i} = x_{t_{i-1}} + \alpha x_{t_{i-1}}\Delta t + \sigma \varepsilon_{t_i} \text{ where } \varepsilon_{t_i} \sim N(\varepsilon_{t_i}; 0, \Delta t) \quad (3.50)$$

Next we replace the discrete real time points $0 = t_0, \dots, t_n = s$ by integer indices $t = 1, 2, \dots, T := n + 1$ resulting in

$$x_t = x_{t-1} + \alpha x_{t-1}\Delta t + \sigma \varepsilon_t \text{ where } \varepsilon_t \sim N(\varepsilon_t; 0, \Delta t) \quad (3.51)$$

Using $\mathbb{V}(aX) = a^2\mathbb{V}(X)$, we can rewrite the above as

$$x_t = (1 + \alpha\Delta t)x_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim N(\varepsilon_t; 0, \sigma^2\Delta t) \quad (3.52)$$

Finally, we define

$$a := (1 + \alpha\Delta t) \text{ and } \sigma_x^2 := \sigma^2\Delta t \quad (3.53)$$

and obtain the AR(1) form

$$x_t = ax_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \sim N(\varepsilon_t; 0, \sigma_x^2) \quad (3.54)$$

4 An Introduction to Variational Calculus

The basic problem of variational calculus is the optimization, i.e. minimization or maximization, of functionals with respect to their arguments (informally, a functional is a “function of a function”. It takes a function as its argument and maps it onto a real-valued scalar). In other words, given a functional $\mathcal{F}$ defined on a function space V , the aim of variational calculus is to find an element y of V for which $\mathcal{F}$ assumes an extremal value. This basic problem can be understood in analogy to the maximization/minimization of a function defined on a standard vector space, such as $\mathbb{R}$ or $\mathbb{R}^n$. In fact, the basic approach of solving the variational optimization problem shows many intuitive similarities with approaches from ordinary calculus (Figure 3, panels A and B). In ordinary calculus, the argument x that maximizes a given function f is usually found by first identifying the stationary points of f , i.e. those arguments x , for which the derivative or gradient of f vanishes. This is the necessary condition for an extremal value: If f has a maximum or a minimum in x , then its gradient in x is equal to zero. However, the condition of a vanishing gradient is not sufficient for a (local) extremal value: The gradient may also vanish for an inflection point or for constant parts of f . Similar considerations apply for the calculus of variations. The formal treatment of necessary and sufficient conditions of functionals is a central topic of functional analysis, just as real analysis represents a formal treatment of ordinary calculus. In the current tutorial, functional analysis is eschewed and only the necessary condition for extremals of functionals will be considered. This approach is somewhat justified by the applied character of this tutorial and the intuition that the functionals considered are usually convex. For a convex functional the necessary condition for an extremum is also sufficient for a global minimum as it is for convex functions (Boyd & Vandenberghe, 2004). The basic problem of variational calculus may thus be formalized as follows: Let V be a vector space of functions, for example, the space of all real-valued differentiable functions on the real line $Y := \{y \in \mathcal{C}^1(\mathbb{R})\}$. Further, let $\mathcal{F}$ be a real-valued functional on Y , i.e., let $\mathcal{F}$ be a mapping that takes an element y of Y as input and maps it onto the real line, formally $\mathcal{F}: Y \rightarrow \mathbb{R}$. Then, the basic problem of variational calculus may be stated as finding $y^* \in Y$ for which $\mathcal{F}$ assumes an extremum (a minimum or a maximum):

$$y^* := \arg \min_{y \in V} \mathcal{F}(y) = \arg \max_{y \in V} (-\mathcal{F}(y)) \quad (4.1)$$

A necessary condition for such an extremal argument (and, by appealing to the intuition from ordinary calculus, an equation of the type $f'(x) = 0$ which can be solved for the respective argument x) can be established by introducing the so-called “Gateaux derivative” of $\mathcal{F}$: Let y^* be a minimum of $\mathcal{F}$ and let $v \in Y$ be another function in Y . Because Y is a function space, $y^* + \varepsilon v$ for $\varepsilon \in \mathbb{R}$ is also an element of Y . Moreover, because y^* is a minimum of $\mathcal{F}$

$$\mathcal{F}(y^* + \varepsilon v) \geq \mathcal{F}(y^*) \quad (4.2)$$

This consideration allows for defining a function of ε as follows:

$$h_{y^*,v}: \mathbb{R} \rightarrow \mathbb{R}, \quad \varepsilon \mapsto h_{y^*,v}(\varepsilon) := \mathcal{F}(y^* + \varepsilon v) \quad (4.3)$$

By definition, this function has a minimum in $\varepsilon = 0$ as $\mathcal{F}(y^* + 0 \cdot v) = \mathcal{F}(y^*) \leq \mathcal{F}(y^* + \varepsilon v)$. Thus, if $h_{y^*,v}$ is differentiable, its derivative with respect to ε vanishes in $\varepsilon = 0$, in other words,

$$\frac{d}{d\varepsilon} h_{y^*,v}(\varepsilon)|_{\varepsilon=0} = \lim_{\varepsilon \rightarrow 0} \frac{\mathcal{F}(y^* + \varepsilon v) - \mathcal{F}(y^*)}{\varepsilon} = 0 \quad (4.4)$$

This derivative of $h_{y^*,v}$ in $\varepsilon = 0$ (if defined) is referred to as the Gateaux derivative of $\mathcal{F}$ and is usually denoted as $\delta\mathcal{F}(y^*, v)$. One thus obtains

$$\delta\mathcal{F}(y^*, v) := \lim_{\varepsilon \rightarrow 0} \frac{\mathcal{F}(y^* + \varepsilon v) - \mathcal{F}(y^*)}{\varepsilon} = 0 \Leftrightarrow \frac{d}{d\varepsilon} h_{y^*,v}(\varepsilon)|_{\varepsilon=0} = 0 \quad (4.5)$$

A necessary condition for an extremal value of the functional $\mathcal{F}$ can thus be expressed as follows: Let y^* be an extremal value of $\mathcal{F}$, i.e., $y^* := \arg \min_{y \in V} \mathcal{F}(y)$. Then it follows that

$$\delta \mathcal{F}(y^*, v) = 0 \quad (4.6)$$

for all $v \in Y$ given that $\delta \mathcal{F}(y^*, v)$ exists. This condition thus provides an approach for determining an extremal argument of a functional: after determining $\delta \mathcal{F}(y, v)$ for a given functional $\mathcal{F}$, set $\delta \mathcal{F}(y, v) = 0$ and solve for y . This approach yields the “stationary point(s)” of $\mathcal{F}$, which can then be investigated with respect to their extremal properties. Further, in the case of a convex functional $\mathcal{F}$, these solutions already constitute the minimum argument(s).

4.1 The Variational Problem with Fixed Endpoints

Above, a necessary condition for an extremal argument y^* of a functional $\mathcal{F}$ was introduced within a very general context. Here, this general approach is made concrete for a specific class of functionals (subsuming, in a sense, the variational free energy considered in Section 3 of the main tutorial) and is demonstrated by means of an example. Specifically, by limiting the function space to real-valued functions on an interval $[x_0, x_1]$ and specifying the class of functionals as the integrals of these functions, one arrives at the Euler-Lagrange equation for functional optimization. As will be shown below, the Euler-Lagrange equation defines the extremal argument y^* of a functional $\mathcal{F}$ by means of a second-order partial differential equation, which can be solved using methods from ordinary calculus.

The variational fixed endpoint problem may be stated as follows: let $(x_0, y_0), (x_1, y_1)$ be elements of $\mathbb{R}^2$ and Y the vector space of all twice-differentiable functions y on (x_0, x_1) which fulfill the endpoint conditions $y(x_0) = y_0$ and $y(x_1) = y_1$, i.e.,

$$Y := \{y \in \mathcal{C}^2[x_0, x_1] | y(x_0) = y_0, y(x_1) = y_1\} \quad (4.7)$$

Further, let L be a twice partially differentiable, real-valued function defined on the set $\mathbb{R} \times \mathbb{R} \times [x_0, x_1]$:

$$L: \mathbb{R} \times \mathbb{R} \times [x_0, x_1] \rightarrow \mathbb{R} \quad (4.8)$$

Then, the variational fixed endpoint problem consist of finding a $y^* \in Y$ such that the functional $\mathcal{F}$ defined by

$$\mathcal{F}: V \rightarrow \mathbb{R}, y \mapsto \mathcal{F}(y) := \int_{x_0}^{x_1} L(y(x), y'(x), x) dx \quad (4.9)$$

becomes extremal, in other words, finding

$$y^* := \arg \min_{y \in V} \int_{x_0}^{x_1} L(y, y', x) dx \quad (4.10)$$

Note that, y and y' refer to both the functions $y, y' \in V$ and the real coordinates $y(x), y'(x) \in \mathbb{R}$ of the function L for $x \in \mathbb{R}$. Specifying the Gateaux derivative of $\mathcal{F}$ as in (4.5) yields the following necessary condition for an extremal value in the current fixed endpoint case

$$\delta \mathcal{F}(y^*, v) = \frac{d}{d\varepsilon} \int_{x_0}^{x_1} L(y^* + \varepsilon v, y^{*'} + \varepsilon v', x) dx|_{\varepsilon=0} = 0 \quad (4.11)$$

with $v \in V$ where, to respect the boundary conditions,

$$V := \{v \in \mathcal{C}^2[x_0, x_1] | v(x_0) = 0, v(x_1) = 0\}$$

so that $y + v \in Y$ for arbitrary y and v .

It can be shown that the derivative of the integral in (4.11) may be reformulated such that the necessary condition above may equivalently be expressed as (see below for details; for simplicity of notation, the asterisk denoting the extremal argument is suppressed in the following equations, which also emphasizes the notion of y being a variable which is solved for)

$$\delta\mathcal{F}(y, v) = 0 \Leftrightarrow \frac{\partial}{\partial y} L(y, y', x) - \frac{d}{dx} \frac{\partial}{\partial y'} L(y, y', x) = 0 \quad (4.12)$$

The equation on the right-hand side of the expression above is referred to as the Euler-Lagrange equation and represents an (implicit) second-order partial differential equations for y . This can be seen by evaluating the total differential of its second term resulting in (see below for details)

$$\frac{\partial}{\partial y} L(y, y', x) = \frac{d}{dx} \frac{\partial}{\partial y'} L(y, y', x) \Leftrightarrow \frac{\partial}{\partial y} L(y, y', x) = \frac{\partial^2}{\partial y \partial y'} L(y, y', x) y' + \frac{\partial^2}{\partial y' \partial y'} L(y, y', x) y'' + \frac{\partial^2}{\partial x \partial y'} L(y, y', x) \quad (4.13)$$

This implicit partial differential equation for y may be made explicit by dividing the second partial derivative of the Lagrange function with respect to y' and evaluated using standard methods from ordinary calculus. The resulting $y \in V$ is a stationary point of $\mathcal{F}$ and a candidate for an extremal argument. Again, if $\mathcal{F}$ is convex, it is also already known to be an extremal argument. In the following example, the calculus of partial differential equations is eschewed, as the functional considered is an explicit function of y' only.

◦ Derivation of equations (4.12) and (4.13)

In this section, we derive the Euler-Lagrange equation

$$\frac{\partial}{\partial y} L(y, y', x) - \frac{d}{dx} \frac{\partial}{\partial y'} L(y, y', x) = 0 \quad (4.14)$$

based on the necessary condition for a minimum of

$$\mathcal{F}(y) := \int_{x_0}^{x_1} L(y(x), y'(x), x) dx \quad (4.15)$$

given by

$$\delta\mathcal{F}(y, v) = \frac{d}{d\varepsilon} \int_{x_0}^{x_1} L(y + \varepsilon v, y' + \varepsilon v', x) dx|_{\varepsilon=0} = 0 \quad (4.16)$$

To this end, we first note the definition of the total differential for a function $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ and $\tilde{y} := (y, y', x)^T, \tilde{v} := (v, v', 0)^T, \varepsilon \in \mathbb{R}$ given by

$$\frac{d}{d\varepsilon} f(\tilde{y} + \varepsilon \tilde{v})|_{\varepsilon=0} = \sum_{i=1}^n \frac{\partial}{\partial y_i} f(\tilde{y}) \cdot v_i \quad (4.17)$$

Because the Lagrange function L is differentiable by definition, we thus obtain

$$\begin{aligned} \frac{d}{d\varepsilon} \int_{x_0}^{x_1} L(y + \varepsilon v, y' + \varepsilon v', x) dx|_{\varepsilon=0} &= \int_{x_0}^{x_1} \frac{d}{d\varepsilon} (L(y + \varepsilon v, y' + \varepsilon v', x))|_{\varepsilon=0} dx \\ &= \int_{x_0}^{x_1} \frac{\partial}{\partial y} L(y, y', x) \cdot v + \frac{\partial}{\partial y'} L(y, y', x) \cdot v' + \frac{\partial}{\partial x} L(y, y', x) \cdot 0 dx \\ &= \int_{x_0}^{x_1} \frac{\partial}{\partial y} L(y, y', x) v + \frac{\partial}{\partial y'} L(y, y', x) v' dx \\ &= \int_{x_0}^{x_1} \frac{\partial}{\partial y} L(y, y', x) v dx + \int_{x_0}^{x_1} \frac{\partial}{\partial y'} L(y, y', x) v' dx \end{aligned} \quad (4.18)$$

The second integral term is next integrated using partial integration. In general we have

$$\int_a^b f' g = f g|_a^b - \int_a^b f g' \quad (4.19)$$

Here, we have

$$a = x_1, \quad b = x_1, \quad g = \frac{\partial}{\partial y'} L, \quad g' = \frac{d}{dx} \frac{\partial}{\partial y'} L, \quad f = v, \quad f' = v' \quad (4.20)$$

and thus obtain

$$\int_{x_0}^{x_1} \frac{\partial}{\partial y'} L \cdot v' dx = \frac{\partial}{\partial y'} L \cdot v|_{x_0}^{x_1} - \int_{x_0}^{x_1} v \cdot \frac{d}{dx} \frac{\partial}{\partial y'} L \quad (4.21)$$

Written in full, the above yields

$$\begin{aligned} & \int_{x_0}^{x_1} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) v' dx \\ &= \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) v(x) \Big|_{x_0}^{x_1} - \int_{x_0}^{x_1} \frac{d}{dx} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) v(x) dx \\ &= \frac{\partial}{\partial y'(x)} L(y(x_1), y'(x_1), x_1) v(x_1) - \frac{\partial}{\partial y'(x)} L(y(x_0), y'(x_0), x_0) v(x_0) - \int_{x_0}^{x_1} \frac{d}{dx} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) v(x) dx \end{aligned} \quad (4.22)$$

As we are considering only v for which $v(x_0) = v(x_1) = 0$, we have

$$\int_{x_0}^{x_1} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) v' dx = - \int_{x_0}^{x_1} \frac{d}{dx} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) v(x) dx \quad (4.23)$$

Substitution into the equation above results in

$$\begin{aligned} & \delta \mathcal{F}(y, v) = 0 \\ & \Leftrightarrow \int_{x_0}^{x_1} \frac{\partial}{\partial y(x)} L(y(x), y'(x), x) v(x) dx - \int_{x_0}^{x_1} \frac{d}{dx} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) v(x) dx = 0 \\ & \Leftrightarrow \int_{x_0}^{x_1} \left(\frac{\partial}{\partial y} L(y(x), y'(x), x) - \frac{d}{dx} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) \right) v(x) dx = 0 \end{aligned} \quad (4.24)$$

We now make use of the following Lagrange-Lemma, which we state without proof:

Lagrange-Lemma

Let f be a continuous real-valued function on the interval $[x_0, x_1]$ and let

$$\int_{x_0}^{x_1} f(x) v(x) dx = 0$$

for all continuously differentiable, real-valued functions v that fulfill $v(x_0) = v(x_1) = 0$. Then it follows that $f(x) \equiv 0$ on $[x_0, x_1]$.

With this Lemma, we thus obtain the Euler-Lagrange equation

$$\delta \mathcal{F}(y, v) = 0 \Leftrightarrow \frac{\partial}{\partial y(x)} L(y(x), y'(x), x) - \frac{d}{dx} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) = 0 \quad (4.25)$$

Finally, we note that the Euler-Lagrange equation constitutes a second order partial differential equation for the function y , because writing out the differential of the second term on its left hand side yields

$$\frac{d}{dx} \frac{\partial}{\partial y'(x)} L(y(x), y'(x), x) = \frac{\partial^2}{\partial y(x) \partial y'(x)} L(y(x), y'(x), x) y'(x) + \frac{\partial^2}{\partial y'(x) \partial y'(x)} L(y(x), y'(x), x) y''(x) + \frac{\partial^2}{\partial y'(x) \partial x} L(y(x), y'(x), x) \quad (4.26)$$

Written in full, the Euler-Lagrange equation in its general form thus corresponds to

$$\frac{\partial}{\partial y(x)} L(y(x), y'(x), x) - \frac{\partial^2}{\partial y(x) \partial y'(x)} L(y(x), y'(x), x) y'(x) - \frac{\partial^2}{\partial y'(x) \partial y'(x)} L(y(x), y'(x), x) y''(x) - \frac{\partial^2}{\partial y'(x) \partial x} L(y(x), y'(x), x) = 0 \quad (4.27)$$

□

Example: The Shortest Distance Between Two Points in $\mathbb{R}^2$

As an example, we consider the following problem: Let $(a, \alpha), (b, \beta) \in \mathbb{R}^2, a < b$. The vector space Y may be defined as

$$Y := \{y \in \mathcal{C}^1[a, b] | y(a) = \alpha, y(b) = \beta\} \quad (4.28)$$

and the aim is to find a $y \in V$ for which the functional

$$\mathcal{F}: Y \rightarrow \mathbb{R}, y \mapsto \mathcal{F}(y) := \int_a^b \sqrt{1 + y'(x)^2} dx \quad (4.29)$$

becomes minimal. The specific form of the functional in this example is motivated by the fact that, for parameterized curves in $\mathbb{R}^2$ (i.e., functions y of the type $f: \mathbb{R} \rightarrow \mathbb{R}^2$), the arc length is defined as

$$l: V \rightarrow \mathbb{R}, l \mapsto l(f) := \int_a^b \|f'(x)\|_2 dx \quad (4.30)$$

and the function y considered here represents the “function form” of such a curve, i.e., $f(x) = (x, y(x))^T$ for which $f'(x)$ evaluates to $(1, y'(x))^T$. Therefore, in the special case of the example considered here, the Lagrange function is given by

$$L: \mathbb{R} \rightarrow \mathbb{R}, y' \mapsto L(y') := \sqrt{1 + y'^2} \quad (4.31)$$

L is thus not explicitly dependent on y and x , the partial derivatives of the Lagrange function with respect to y and x evaluate to zero, and the Euler-Lagrange equation simplifies to

$$\frac{\partial}{\partial y} L(y') = \frac{d}{dx} \frac{\partial}{\partial y'} L(y') \Leftrightarrow \frac{d}{dx} \frac{\partial}{\partial y'} L(y') = 0 \quad (4.32)$$

The right-hand side of the equivalence relation (4.32) states that the total derivative of $\frac{\partial}{\partial y'} L(y')$ is zero and the function $\frac{\partial}{\partial y'} L(y')$ thus constant. One hence obtains an ordinary differential equation for y given by

$$\frac{\partial}{\partial y'} L(y') = c \Rightarrow y'(x) = \sqrt{\frac{c^2}{1 - c^2}} \quad (4.33)$$

for a given constant $c \in \mathbb{R}$ (see below for details). As the derivative of y is thus constant, y must be a linear-affine function of the type

$$y(x) = sx + r, s, r \in \mathbb{R} \quad (4.34)$$

Based on the fixed endpoints $y(a) = \alpha, y(b) = \beta$, the constants s and r may be evaluated, resulting in the solution

$$y^*(t) := \arg \min_{y \in V} \int_a^b \sqrt{1 + y'(x)^2} dx = \frac{\beta - \alpha}{b - a} (x - a) + \alpha \quad (4.35)$$

The application of the variational method hence verifies the intuition that the functional form of the curve representing the shortest distance between two points in the plane is a straight line (Figure 2 panels C and D).

○ Derivation of equation (4.33)

We first evaluate the partial derivative $\frac{\partial}{\partial y'} L(y')$ for $L(y') := \sqrt{1 + y'^2}$:

$$\frac{\partial}{\partial y'} L(y') = \frac{\partial}{\partial y'} (1 + y'^2)^{\frac{1}{2}} = \frac{1}{2} (1 + y'^2)^{-\frac{1}{2}} \cdot 2 \cdot y' = \frac{y'}{\sqrt{1 + y'^2}} \quad (4.36)$$

The Euler Lagrange Equation

$$\frac{d}{dx} \left(\frac{y'}{\sqrt{1 + y'^2}} \right) = 0 \quad (4.37)$$

thus implies that

$$\frac{y'}{\sqrt{1 + y'^2}} = c \quad (4.38)$$

for a real constant $c \in \mathbb{R}$. Reformulating then leads to the insight that the derivative of y w.r.t. x is constant:

$$y'^2 = c^2 (1 + y'^2) \Leftrightarrow y'^2 = c^2 + c^2 y'^2 \Leftrightarrow y'^2 - c^2 y'^2 = c^2 \Leftrightarrow y'^2 (1 - c^2) = c^2 \Leftrightarrow y' = \sqrt{\frac{c^2}{1 - c^2}} \quad (4.39)$$

The Isoperimetric Variational Problem

In addition to the conditions of fixed endpoints, the extremal function maximizing/minimizing a given functional in the context of isoperimetric problems have to fulfill additional integral conditions. In general, the isoperimetric variational problem may be stated as follows: let $(x_0, y_0), (x_1, y_1)$ be elements of $\mathbb{R}^2$ and Y the vector space of all twice-differentiable functions y on (x_0, x_1) which fulfill the endpoint conditions $y(x_0) = y_0$ and $y(x_1) = y_1$. Further, let L and G be twice partially differentiable, real-valued functions defined on the set $\mathbb{R} \times \mathbb{R} \times [x_0, x_1]$. Then, the isoperimetric variational problem comprises finding a $y^* \in Y$ such that the functional $\mathcal{F}$ defined by

$$\mathcal{F}: Y \rightarrow \mathbb{R}, y \mapsto \mathcal{F}(y) := \int_{x_0}^{x_1} L(y(x), y'(x), x) dx \quad (4.40)$$

becomes extremal, whilst fulfilling the integral side-condition

$$I: Y \rightarrow \mathbb{R}, y \mapsto I(y) := \int_{x_0}^{x_1} G(y(x), y'(x), x) dx = 0 \quad (4.41)$$

The formal development of an approach for the solution of the isoperimetric variational problem proving the existence of Lagrangian multipliers falls into the realm of functional analysis and is eschewed here. Nevertheless, assuming the existence of both an extremal argument y^* and suitable Lagrangian multipliers λ , the Lagrange Lemma introduced in below provides the following recipe for solving an isoperimetric variational problem:

(1) consider the Lagrange function $F: \mathbb{R} \times \mathbb{R} \times [x_0, x_1] \rightarrow \mathbb{R}$, defined by

$$F(y(x), y'(x), x) := L(y(x), y'(x), x) + \lambda G(y(x), y'(x), x) \quad (4.42)$$

(2) instead of minimizing $\mathcal{F}$, minimize the extended functional

$$\tilde{\mathcal{F}} := \int_{x_0}^{x_1} F(y(x), y'(x), x) dx \quad (4.43)$$

using the fixed endpoint approach developed in Section 3.2 to obtain the extremal argument y^* (usually by means of the corresponding Euler-Lagrange equations) and

(3) determine λ so that y^* solves the side-condition $\int_{x_0}^{x_1} G(y(x), y'(x), x) dx = 0$.

If this approach is possible, the Lagrange Lemma guarantees that the resulting y^* fulfills the isoperimetric side-condition while minimizing the initial functional. This approach is demonstrated using a classic example below.

It should be noted that the functions maximizing the functional of interest of the VML and VB frameworks are usually defined on the entire real line, rather than on a real interval with fixed endpoint conditions, as considered here. For simplicity, however, a formal treatment of the ensuing improper integrals is eschewed, and the methods developed in the current section are transferred to the problems considered in the next section by the intuition that the probability density functions of interest are usually Gaussian for which $f(x) \rightarrow 0$ in the limits of $x \rightarrow \infty$ and $x \rightarrow -\infty$.

The Lagrange Lemma for the isoperimetric variational problem

If the existence of a minimum $\hat{y}$ and of the corresponding Lagrangian multiplier λ is postulated, the following Lemma can readily be proven, and provides an ansatz for the solution of the isoperimetric variational problem:

Lagrange-Lemma for the isoperimetric variational problem

Let M be an arbitrary set and $J: M \rightarrow \mathbb{R}$ a function to be minimized on the restriction set

$$R := \{y \in M | g(y) = 0\}$$

where $g: M \rightarrow \mathbb{R}$ defines a side condition. Further, let $\lambda \in \mathbb{R}$ and $\hat{y} \in R$ such that $\hat{y}$ is a minimum of

$$\tilde{\mathcal{F}} := J + \lambda g : M \rightarrow \mathbb{R}$$

Then $\hat{y}$ is a Minimum of J on R .

The Lagrange Lemma for the isoperimetric variational problem can be verified by considering the following:

$$J(\hat{y}) = J(\hat{y}) + 0 = J(\hat{y}) + \lambda g(\hat{y}) \leq J(y) + \lambda g(y) = J(y) \quad (4.44)$$

Thus, if the problem of minimizing the extended functional

$$\tilde{\mathcal{F}} := \int_{x_0}^{x_1} F(y(x), y'(x), x) dx \quad (4.46)$$

where

$$F(y(x), y'(x), x) := L(y(x), y'(x), x) + \lambda G(y(x), y'(x), x) \quad (4.47)$$

as specified in equations (4.40) and (4.41) can be solved, its solution provides a minimum of $\mathcal{F}$ which fulfills the side condition specified by I .

Example: The Problem of Dido

The problem of Dido, to fence in a maximum area with a boundary of given length, may be stated as an isoperimetric variational problem as follows: Consider the vector space

$$Y := \{y \in \mathcal{C}^2[a, b] | y(a) = 0 = y(b)\} \quad (4.48)$$

The aim is to find a $y \in Y$, which minimizes the functional

$$\mathcal{F}: Y \rightarrow \mathbb{R}, y \mapsto \mathcal{F}(y) := - \int_a^b y(x) dx \quad (4.49)$$

subject to the constraint (side-condition) that

$$g(y) = 0, \text{ where } g: Y \rightarrow \mathbb{R}, y \mapsto g(y) := \int_a^b \sqrt{1 + y'(x)^2} dx - l \quad (4.50)$$

for some constant $l \in \mathbb{R}$. Intuitively, this problem may be stated as maximizing the area between the graph of y and the x -axis using a y of a given constant arc length l (Figure 2, panels C and D). Applying the recipe for the solution of the isoperimetric variational problem discussed above, one obtains the following extended functional (see below for details):

$$\tilde{\mathcal{F}} := - \int_a^b y(x) dx + \lambda \left(\int_a^b \sqrt{1 + y'(x)^2} dx - l \right) = \int_a^b \left(-y(x) + \lambda \sqrt{1 + y'(x)^2} - \frac{\lambda l}{b-a} \right) dx \quad (4.51)$$

where the Lagrange function is not explicitly dependent on x and is given by

$$F: \mathbb{R} \times \mathbb{R}, (y, y') \mapsto L(y, y') = -y + \lambda \sqrt{1 + y'^2} - \frac{\lambda l}{b-a} \quad (4.52)$$

For a Lagrange function that explicitly only depends on y and y' , the Euler-Lagrange equation simplifies to (see below for details)

$$F(y, y') - \frac{\partial}{\partial y'} L(y, y') y' = B \quad (4.53)$$

where $B \in \mathbb{R}$ is a constant. Evaluation of the Euler-Lagrange equation for the problem of Dido results in the following first-order ordinary differential equation for y (see below for details):

$$\sqrt{1 + y'^2} \left(-y - \frac{\lambda l}{b-a} - B \right) + \lambda = 0 \quad (4.54)$$

A solution for this differential equation is given by

$$y(x) = C + \sqrt{\lambda^2 - (x - D)^2} \quad (4.55)$$

where $C := -B - \frac{\lambda l}{b-a}$ and $D \in \mathbb{R}$ is an additional integration constant (for a derivation of this solution, see (Richter, Mathias, o. J.)). Taking the square of the equation above then results in

$$(y(x) - C)^2 + (x - D)^2 = \lambda^2 \quad (4.56)$$

which is recognized as the functional equation for a circle with the center at $(D, C) \in \mathbb{R}^2$ and radius λ . Thus the solution to the problem of Dido is a circular arc, the exact form of which depends on the relationship between the endpoints a, b and the given arc length l (Figure 2, panels C and D). Consideration of the endpoints yields $D = \frac{a+b}{2}$ while analysis of the side-condition allows for inferring λ and subsequently C which can lie above or below the abscissa depending on the parameter.

○ Derivation of equation (4.53)

Provided that L is not an explicit function of x , and thus $\frac{\partial}{\partial x} L = \frac{\partial^2}{\partial y' \partial x} L = 0$ the Euler-Lagrange equation (4.25) yields

$$\begin{aligned}
 & \frac{\partial}{\partial y} L - \frac{d}{dx} \frac{\partial}{\partial y'} L = 0 \\
 & \Leftrightarrow y' \left(\frac{\partial}{\partial y} L - \frac{d}{dx} \frac{\partial}{\partial y'} L \right) = 0 \\
 & \Leftrightarrow y' \frac{\partial}{\partial y} L - \left(y' \frac{\partial^2}{\partial x \partial y'} L + y'^2 \frac{\partial^2}{\partial y \partial y'} L + y' y'' \frac{\partial^2}{\partial y'^2} L \right) = 0 \\
 & \Leftrightarrow \left(\frac{\partial}{\partial x} L + y' \frac{\partial}{\partial y} L + y'' \frac{\partial}{\partial y'} L \right) - \left(y'' \frac{\partial}{\partial y'} L + y' \frac{\partial^2}{\partial x \partial y'} L + y'^2 \frac{\partial^2}{\partial y \partial y'} L + y' y'' \frac{\partial^2}{\partial y'^2} L \right) = 0 \\
 & \Leftrightarrow \frac{d}{dx} L - \frac{d}{dx} \left(y' \frac{\partial}{\partial y'} L \right) = 0 \\
 & \Leftrightarrow \frac{d}{dx} \left(L - y' \frac{\partial}{\partial y'} L \right) = 0
 \end{aligned} \tag{4.57}$$

Therefore

$$L(y, y') - \frac{\partial}{\partial y'} L(y, y') y' = B \tag{4.58}$$

○ Derivation of equation (4.54)

We apply

$$F(y, y') - \frac{\partial}{\partial y'} F(y, y') y' = B \tag{4.60}$$

to the Lagrange function given in (4.52)

$$F(y, y') := -y + \lambda \sqrt{1 + y'^2} - \frac{\lambda l}{b-a} \tag{4.61}$$

and obtain

$$\begin{aligned}
 & -y + \lambda \sqrt{1 + y'^2} - \frac{\lambda l}{b-a} - y' \frac{\partial}{\partial y'} \left(-y + \lambda \sqrt{1 + y'^2} - \frac{\lambda l}{b-a} \right) = B \\
 & \Leftrightarrow -y + \lambda \sqrt{1 + y'^2} - \frac{\lambda l}{b-a} - y' \frac{\lambda y'}{\sqrt{1 + y'^2}} = B
 \end{aligned} \tag{4.62}$$

Multiplying both sides by $\sqrt{1 + y'^2}$ yields

$$\begin{aligned}
 & -y \sqrt{1 + y'^2} + \lambda (1 + y'^2) - \frac{\lambda l}{b-a} \sqrt{1 + y'^2} - \lambda y'^2 = B \sqrt{1 + y'^2} \\
 & \Leftrightarrow -y \sqrt{1 + y'^2} + \lambda + \lambda y'^2 - \frac{\lambda l}{b-a} \sqrt{1 + y'^2} - \lambda y'^2 - B \sqrt{1 + y'^2} = 0 \\
 & \Leftrightarrow \sqrt{1 + y'^2} \left(-y - \frac{\lambda l}{b-a} - B \right) + \lambda = 0
 \end{aligned} \tag{4.63}$$

□

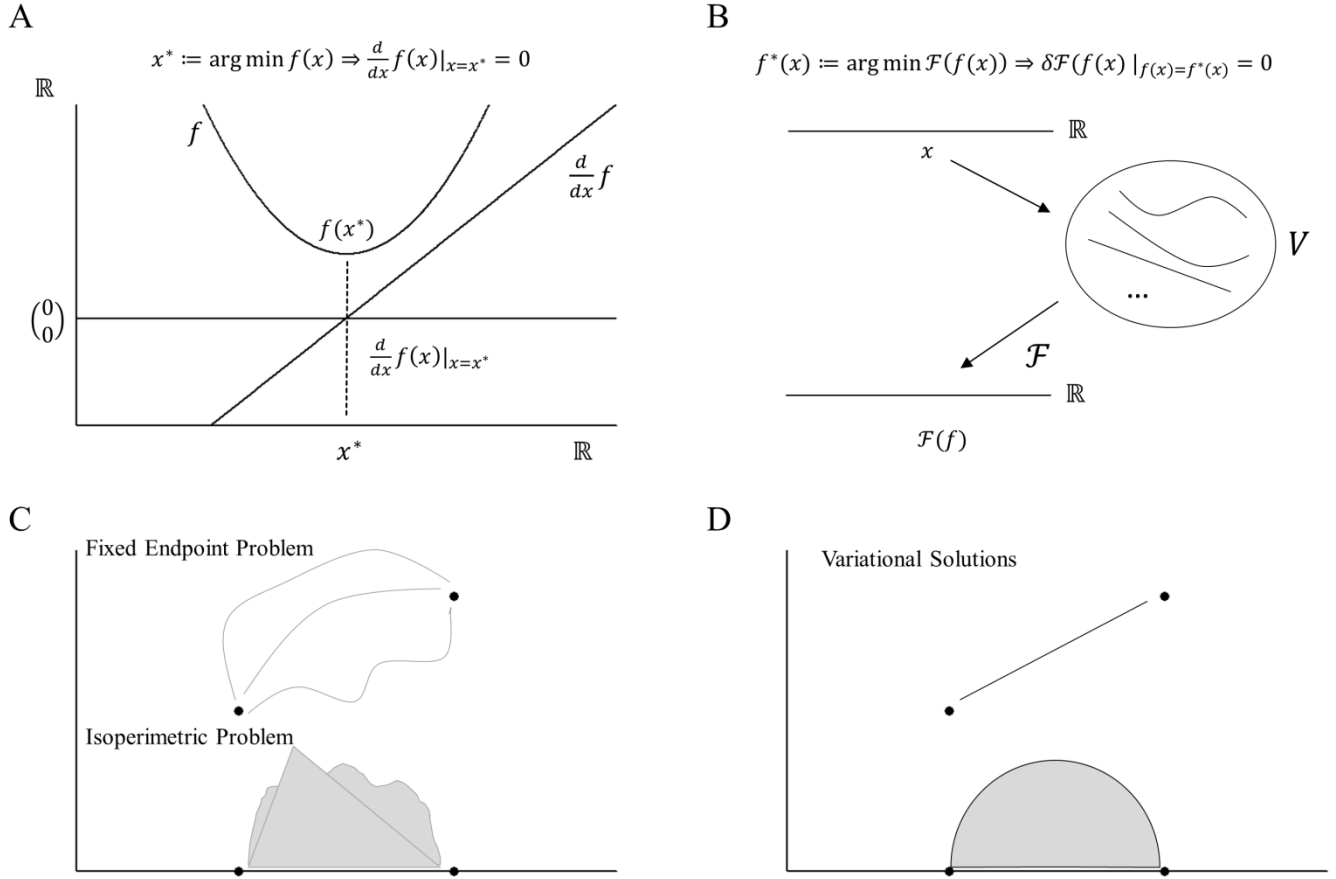


Figure 2. An intuitive notion of the calculus of variations. Figure 2A depicts a standard calculus analogy to the extremal problem of variational calculus. To find the extremal points x^* from a given real-valued function f , standard calculus usually proceeds by finding its stationary points. The stationary points of a function are those arguments of the function for which the derivative $\frac{d}{dx} f$ of the function vanishes, i.e., for which $\frac{d}{dx} f(x)|_{x=x^*} = 0$ holds. Figure 2B depicts a conceptual overview of the variational calculus framework: central to variational calculus is the problem of identifying a (real-valued) function f in a function space V , which extremizes a given functional $\mathcal{F}$. The functionals considered in this review (especially the variational free energy) map functions (in the context of the tutorial probability density functions) onto the real line. Candidate functions are found by capitalizing on the necessary condition for an extremal, i.e., a vanishing Gateaux derivative $\delta \mathcal{F}(f(x))|_{f(x)=f^*(x)} = 0$ and solving for $f^*(x)$. Figures 2C and 2D sketch the variational problems and the solutions to them considered as examples in Section 3. The fixed endpoint problem can be conceived as the problem to find the curve (i.e., a mapping $f: \mathbb{R} \rightarrow \mathbb{R}^n$, here, $n = 2$) of minimal length connecting two given points in $\mathbb{R}^2$. The application of the variational calculus method allows for identifying the straight line connecting both points as the optimal function (curve). The isoperimetric problem (here, with fixed endpoints) represents a variational problem with additional integral side-conditions. The traditional example for this is the problem of Dido, which lies in finding the function of given arc length that maximizes the area between the function and the ordinate. Using the calculus of variations and a Lagrange multiplier approach, it can be shown that the solutions are circular arcs whose exact form depends on the predetermined arc length. Maximization of the variational free energy in the VB framework may also be viewed as an isoperimetric problem. Here, the functions maximizing the variational free energy need to fulfill the integral side-condition of being probability density functions on the real line, i.e., their integral over the real line should equal 1.

Derivation of the VB inference theorem for mean-field approximation

In the VB framework the aim is to approximate the log marginal likelihood $\ln p(y)$ by iteratively maximizing its lower bound $\mathcal{F}(q(\vartheta_s), q(\vartheta_{\setminus s}))$ with respect to the arguments $q(\vartheta_s)$ and $q(\vartheta_{\setminus s})$. For iterations $i = 1, 2, \dots$, during the first maximization of finding $q^{(i+1)}(\vartheta_s)$, $q^{(i)}(\vartheta_{\setminus s})$ is treated as a constant, while during the second maximization of finding $q^{(i+1)}(\vartheta_{\setminus s})$, $q^{(i+1)}(\vartheta_s)$ is treated as a constant. However, as ϑ_s and $\vartheta_{\setminus s}$ may be used interchangeably, we here concern ourselves only with the case of maximizing $\mathcal{F}(q(\vartheta_s), q(\vartheta_{\setminus s}))$ with respect to $q(\vartheta_s)$. To obtain an expression for $q^{(i+1)}(\vartheta_s)$, we thus consider the functional

$$\mathcal{F}(q(\vartheta_s), q^{(i)}(\vartheta_{\setminus s})) = \iint q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s}) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s})} \right) d\vartheta_s d\vartheta_{\setminus s} \quad (4.65)$$

with the constraint

$$\int q(\vartheta_s) d\vartheta_s = 1 \Leftrightarrow \int q(\vartheta_s) d\vartheta_s - 1 = 0 \Leftrightarrow \int q(\vartheta_s) - \delta(\vartheta_s) d\vartheta_s = 0 \quad (4.66)$$

where δ denotes the Dirac delta function which integrates to 1.

The Lagrange function is given in this case as

$$F(q(\vartheta_s), q^{(i)}(\vartheta_{\setminus s})) = \int q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s}) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s})} \right) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda \delta(\vartheta_s) \quad (4.67)$$

and, as in the previous section, the Gateaux derivative $\delta \mathcal{F}(q(x), q^{(i)}(\theta))$ is given by the derivative of F with respect to $q(\vartheta_s)$, as F is not a function of $q'(\vartheta_s)$. One thus obtains with a constant $C \in \mathbb{R}$ (see below for details)

$$\delta \tilde{\mathcal{F}}(q(\vartheta_s), q^{(i)}(\vartheta_{\setminus s})) = \int q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s} - \ln q(\vartheta_s) + C \quad (4.68)$$

◦ Derivation of equation (4.68)

$$\begin{aligned} & \delta \tilde{\mathcal{F}}(q(\vartheta_s), q^{(i)}(\vartheta_{\setminus s})) \\ &= \frac{\partial}{\partial q(\vartheta_s)} \left(\int q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s}) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s})} \right) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda \delta(\vartheta_s) \right) \\ &= \frac{\partial}{\partial q(\vartheta_s)} \left(\int q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s} - \int q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s}) \ln (q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s})) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda \delta(\vartheta_s) \right) \\ &= \frac{\partial}{\partial q(\vartheta_s)} \left(q(\vartheta_s) \int q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s} - \int q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s}) \ln q(\vartheta_s) d\vartheta_{\setminus s} - \int q(\vartheta_s) q^{(i)}(\vartheta_{\setminus s}) \ln q^{(i)}(\vartheta_{\setminus s}) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda \delta(\vartheta_s) \right) \\ &= \frac{\partial}{\partial q(\vartheta_s)} \left(q(\vartheta_s) \int q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s} - q(\vartheta_s) \ln q(\vartheta_s) \int q^{(i)}(\vartheta_{\setminus s}) d\vartheta_{\setminus s} - q(\vartheta_s) \int q^{(i)}(\vartheta_{\setminus s}) \ln q^{(i)}(\vartheta_{\setminus s}) d\vartheta_{\setminus s} + \lambda q(\vartheta_s) - \lambda \delta(\vartheta_s) \right) \\ &= \int q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s} - \ln q(\vartheta_s) - 1 - \int q^{(i)}(\vartheta_{\setminus s}) \ln q^{(i)}(\vartheta_{\setminus s}) d\vartheta_{\setminus s} + \lambda \\ &= \int q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s} - \ln q(\vartheta_s) + C \end{aligned} \quad (4.69)$$

□

Setting the Gateaux derivative (4.68) to zero thus yields

$$\ln q^{(i+1)}(\vartheta_s) := \int q^{(i)}(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s} + C \quad (4.70)$$

Taking the exponential and expressing the multiplicative constant as proportionality leads to the “VB inference theorem for mean-field approximations”

$$q^{(i+1)}(\vartheta_s) \propto \exp\left(\int q(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s}\right) \quad (4.71)$$

5 Details of the univariate Gaussian example

5.1 Introduction

To demonstrate the VB inference method for i.i.d. data we consider the Bayesian estimation of the expectation and variance parameters of a univariate Gaussian (discussed in detail in (Michael Chappell, Adrian Groves and Mark Woolrich, 2008; W. D. Penny, 2000) and visualized in Figure 5 of the tutorial). We thus assume that the data is generated by a univariate Gaussian distribution over an observable variable y_i with true, but unknown, expectation parameter μ and inverse variance (precision) parameter λ , i.e.,

$$p(y_i) := N(y_i; \mu, \lambda^{-1}) \quad (5.1)$$

Further, we assume that a set of N i.i.d. realizations $y_1^*, \dots, y_N^*$ of the observable variable y_i has been obtained. The concatenation of this set is denoted as $y := (y_1, \dots, y_N)^T$. In this scenario, a classical maximum likelihood point estimation approach would result in point estimates for μ and λ^{-1} based on the sample mean $\bar{\mu}$ and the (biased) sample variance $\bar{\sigma}^2$

$$\bar{\mu} = \frac{1}{n} \sum_{n=1}^N y_n \text{ and } \bar{\sigma}^2 = \frac{1}{n} \sum_{n=1}^N (y_n - \bar{\mu})^2 \quad (5.2)$$

respectively. As outlined in the tutorial, based on appropriately chosen prior distributions, the aim of the Bayesian paradigm is to obtain posterior probability distributions, which quantify the remaining uncertainty over the true, but unknown, unobservable variables given the observable variables and an approximation to the model evidence, i.e. the probability of the data given the generative model. Following the convention outlined in the tutorial (Section 3.1) “generative model” refers to a joint distribution over the observable and the unobservable variables. Thus in the current example it constitutes a joint probability distribution over the observable variable y and both unobservable random variables μ and λ , $\lambda > 0$. It is given by

$$p(y, \mu, \lambda) = p(\mu, \lambda) p(y|\mu, \lambda) = p(\mu, \lambda) \prod_{n=1}^N p(y_n|\mu, \lambda) \quad (5.3)$$

A possible choice for a prior joint distribution over the unobservable variables is given by the product of a univariate Gaussian distribution over μ and a Gamma distribution over λ according to

$$p(y, \mu, \lambda) = p(\mu) p(\lambda) p(y|\mu, \lambda) = N(\mu; m_\mu, s_\mu^2) G(\lambda; a_\lambda, b_\lambda) \prod_{n=1}^N N(y_n; \mu, \lambda^{-1}) \quad (5.4)$$

Note that many other priors are conceivable. In fact, a more commonly discussed scenario (Murphy, 2012; Bishop 2006) is the case of a non-independent Gaussian-Gamma prior. With respect to the factorized prior considered here, this Gaussian-Gamma prior has the advantage that it is “closed-under-sampling” or, in other words, belongs to the conjugate-exponential class and allows for the derivation of an exact analytical solution for the form of the posterior distribution. On the other hand, it is not clear in which scenarios a dependency of the prior over the expectation parameter μ on the prior over λ is in fact a reasonable assumption. We here discuss the factorized (independent prior) case, as in our view, it corresponds to the more parsimonious choice of prior, and demonstrates how the variational Bayes approach can be used to derive (variational) posterior distributions in model scenarios where no analytical treatment is possible. To this end, we consider a mean-field approximation to the posterior distribution over the data conditional unobservable variables, i.e., we set

$$p(\mu, \lambda|y) \approx q(\mu) q(\lambda) \quad (5.5)$$

Specifically, we set $q(\mu)$ to a normal distribution over μ with variational parameters m_μ^q and s_μ^{2q} , and we set $q(\lambda)$ to a Gamma distribution over λ with variational parameters a_λ^q, b_λ^q resulting in

$$q(\mu)q(\lambda) = N(x; m_\mu^q, s_\mu^{2q})G(\lambda; a_\lambda^q, b_\lambda^q) \quad (5.6)$$

In (5.6), $\{m_\mu^q, s_\mu^{2q}, a_\lambda^q, b_\lambda^q\}$ represents a set of “variational” parameters. For convenience and to keep the notational overhead at bay, the values of the variational parameters on the i th iteration of the VB algorithm derived below will be denoted by $m_\mu^{(i)}, s_\mu^{2(i)}, a_\lambda^{(i)}$, and $b_\lambda^{(i)}$, dropping the “ q ” superscript. To summarize, we have introduced a set of prior parameters $\{m_\mu, s_\mu^2, a_\lambda, b_\lambda\}$ governing the prior distribution $p(\mu, \lambda)$ of the generative model in (5.3) and a set of variational parameters $\{m_\mu^{(i)}, s_\mu^{2(i)}, a_\lambda^{(i)}, b_\lambda^{(i)}\}$ governing the variational approximation $q(\mu)q(\lambda)$ to the posterior distribution $p(\mu, \lambda|y)$.

5.2 Application of the variational inference for approximate posteriors

As discussed in the tutorial the variational inference theorem for mean-field approximations (cf. equation (17)) states that the variational distribution over the unobservable variable partition ϑ_s is given by

$$q(\vartheta_s) \propto \exp\left(\int q(\vartheta_{\setminus s}) \ln p(y, \vartheta) d\vartheta_{\setminus s}\right) \quad (5.7)$$

For the current example, we have the mean field approximation

$$q(\mu, \lambda) = q(\mu)q(\lambda) \quad (5.8)$$

and thus

$$q(\mu) = C_\mu \cdot \exp\left(\int q(\lambda) \ln p(y, \mu, \lambda) d\lambda\right) \quad (5.9)$$

and

$$q(\lambda) = C_\lambda \cdot \exp\left(\int q(\mu) \ln p(y, \mu, \lambda) d\mu\right) \quad (5.10)$$

where C_μ and C_λ denote proportionality constants that render the proportionality statement in (5.7) equations in (5.9) and (5.10). In the following, we derive an iterative scheme based on the equations above. For this purpose, it is helpful (1.) to denote expectations by the expectation operator

$$\langle f(x) \rangle_{p(x)} := \int p(x) f(x) dx \quad (5.11)$$

because it considerably reduces the visual complexity of the expressions involved, (2.) explicitly denote the iterative nature of the approach by indexing the variational distributions $q(\mu)$ and $q(\lambda)$ as $q^{(i)}(\mu)$ and $q^{(i)}(\lambda)$, which also stresses the fact that in (5.9) and (5.10) above, the left hand side variational distribution refers to the $(i+1)$ th iteration, while the right hand side variational distribution refers to the i th iteration, and (3.), because we are dealing with probability density functions of the exponential family, to log transform the expressions above to simplify proceedings. Upon these notational changes, (5.9) and (5.10) can be re-expressed as

$$\ln q^{(i+1)}(\mu) = \langle \ln p(y, \mu, \lambda) \rangle_{q^{(i)}(\lambda)} + \ln C_\mu \quad (5.12)$$

$$\ln q^{(i+1)}(\lambda) := \langle \ln p(y, \mu, \lambda) \rangle_{q^{(i+1)}(\mu)} + \ln C_\lambda \quad (5.13)$$

In the following two sections, we will derive explicit expressions for the variational parameters $m_\mu^{(i+1)}, s_\mu^{2(i+1)}$ and $a_\lambda^{(i+1)}, b_\lambda^{(i+1)}$ that govern the variational distributions $q^{(i+1)}(\mu)$ and $q^{(i+1)}(\lambda)$, respectively. These expressions are given in terms of the observable data, as well as the preceding variational parameters $m_\mu^{(i)}, s_\mu^{2(i)}, a_\lambda^{(i)}, b_\lambda^{(i)}$ and the prior parameters $m_\mu, s_\mu^2, a_\lambda, b_\lambda$ as will be seen below.

5.3 Variational parameter update-equation for $q(\mu)$

As noted above, the variational inference theorem for mean-field approximations states that the logarithm of the optimal distribution over the unobservable variable μ , $\ln q^{(i+1)}(\mu)$, given a fixed distribution over the parameter variable, $q^{(i)}(\lambda)$, is given by

$$\ln q^{(i+1)}(\mu) = \langle \ln p(y, \mu, \lambda) \rangle_{q^{(i)}(\lambda)} + \ln C_\mu \quad (5.14)$$

which, if we identify the prior distribution $p(\mu)p(\lambda)$ with the initial variational distribution $q^{(0)}(\mu)q^{(0)}(\lambda)$ may be rewritten as (details below)

$$\ln q^{(i+1)}(\mu) = -\frac{1}{2} \langle \lambda \rangle_{q^{(i)}(\lambda)} \sum_{n=1}^N (y_n - \mu)^2 - \frac{1}{2} \frac{(\mu - m_\mu)^2}{s_\mu^2} + \tilde{C}_\mu \quad (5.15)$$

Based on (5.15) and the completing-the-square theorem (Supplement Section 2), one can infer that the optimal variational distribution over the latent variable μ is proportional to a Gaussian

$$q^{(i+1)}(\mu) \propto N\left(\mu; m_\mu^{(i+1)}, s_\mu^{2(i+1)}\right) \quad (5.16)$$

where the updated variational parameters are provided in terms of the data $y_1, \dots, y_N$ and the variational parameters of $q^{(i)}(\mu)$ and $q^{(i)}(\lambda)$ as

$$m_\mu^{(i+1)} := \frac{m_\mu + s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n}{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}} \quad \text{and} \quad s_\mu^{2(i+1)} := \frac{s_\mu^2}{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}} \quad (5.17)$$

○ Derivation of equation (5.15)

For a fixed distribution $q^{(i)}(\lambda)$, we have due to the i.i.d. data assumption and the assumption of factorized priors for $y = y_{1:N}$

$$\begin{aligned} \ln q^{(i+1)}(\mu) &= \langle \ln p(y, \mu, \lambda) \rangle_{q^{(i)}(\lambda)} + \ln C_\mu \\ &= \langle \ln(\prod_{n=1}^N p(y_n, \mu, \lambda)) \rangle_{q^{(i)}(\lambda)} + \ln C_\mu \\ &= \langle \ln(\prod_{n=1}^N p(y_n | \mu, \lambda) p(\mu, \lambda)) \rangle_{q^{(i)}(\lambda)} + \ln C_\mu \\ &= \langle \ln(\prod_{n=1}^N p(y_n | \mu, \lambda) p(\mu) p(\lambda)) \rangle_{q^{(i)}(\lambda)} + \ln C_\mu \\ &= \langle \ln(\prod_{n=1}^N p(y_n | \mu, \lambda)) \rangle_{q^{(i)}(\lambda)} + \langle \ln p(\mu) \rangle_{q^{(i)}(\lambda)} + \langle \ln p(\lambda) \rangle_{q^{(i)}(\lambda)} + \ln C_\mu \end{aligned} \quad (5.18)$$

The variational distribution is initialized with the prior distribution

$$p(\mu)p(\lambda) = q^{(0)}(\mu)q^{(0)}(\lambda) \quad (5.19)$$

Substituting the example specific probability density functions of $p(\mu), p(\lambda), p(y_n|\mu, \lambda), q^{(i)}(\mu)$ and $q^{(i)}(\lambda)$ from (5.4) and (5.6), namely

$$p(\mu) := (2\pi s_\mu^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2s_\mu^2}(\mu - m_\mu)^2\right) \quad (5.20)$$

$$p(\lambda) := \frac{1}{\Gamma(a_\lambda)} \frac{1}{(b_\lambda)^{a_\lambda}} \lambda^{a_\lambda-1} \exp\left(-\frac{\lambda}{b_\lambda}\right) \quad (5.21)$$

$$p(y_n|\mu, \lambda) := \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(y_n - \mu)^2\right) \quad (5.22)$$

$$q^{(i)}(\mu) := (2\pi s_\mu^{2(i)})^{-\frac{1}{2}} \exp\left(-\frac{1}{2s_\mu^{2(i)}}(\mu - m_\mu^{(i)})^2\right) \quad (5.23)$$

$$q^{(i)}(\lambda) := \frac{1}{\Gamma(a_\lambda^{(i)})} \frac{1}{(b_\lambda^{(i)})^{a_\lambda^{(i)}}} \lambda^{a_\lambda^{(i)}-1} \exp\left(-\frac{\lambda}{b_\lambda^{(i)}}\right) \quad (5.24)$$

in (5.18) then yields

$$\begin{aligned} \ln q^{(i+1)}(\mu) &= \langle \ln \left(\prod_{n=1}^N \left(\frac{\lambda}{2\pi} \right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(y_n - \mu)^2\right) \right) \rangle_{q^{(i)}(\lambda)} \\ &\quad + \langle \ln \left((2\pi s_\mu^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2s_\mu^2}(\mu - m_\mu)^2\right) \right) \rangle_{q^{(i)}(\lambda)} \\ &\quad + \langle \ln \left(\frac{1}{\Gamma(a_\lambda)} \frac{1}{(b_\lambda)^{a_\lambda}} \lambda^{a_\lambda-1} \exp\left(-\frac{\lambda}{b_\lambda}\right) \right) \rangle_{q^{(i)}(\lambda)} \\ &\quad + \ln C_\mu \\ &= \langle \frac{N}{2} \ln \lambda - \frac{1}{2} \ln 2\pi - \frac{1}{2} \lambda \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i)}(\lambda)} \\ &\quad + \langle -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln s_\mu^2 - \frac{1}{2} \frac{(\mu - m_\mu)^2}{s_\mu^2} \rangle_{q^{(i)}(\lambda)} \\ &\quad + \langle -\ln(\Gamma(a_\lambda)) - a_\lambda \ln(b_\lambda) + (a_\lambda - 1) \ln \lambda - \frac{\lambda}{b_\lambda} \rangle_{q^{(i)}(\lambda)} \\ &\quad + \ln C_\mu \end{aligned} \quad (5.25)$$

Grouping all constant terms independent of μ (i.e. those terms that do not change if μ changes) with a constant $\tilde{C}_\mu$ and evaluation of the expectation then simplifies the above to

$$\begin{aligned} \ln q^{(i+1)}(\mu) &= -\frac{1}{2} \langle \lambda \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i)}(\lambda)} - \frac{1}{2} \langle \frac{(\mu - m_\mu)^2}{s_\mu^2} \rangle_{q^{(i)}(\lambda)} + \tilde{C}_\mu \\ &= -\frac{1}{2} \langle \lambda \rangle_{q^{(i)}(\lambda)} \sum_{n=1}^N (y_n - \mu)^2 - \frac{1}{2} \frac{(\mu - m_\mu)^2}{s_\mu^2} + \tilde{C}_\mu \end{aligned} \quad (5.26)$$

□

○ Derivation of equations (5.16) and (5.17)

Starting from

$$\ln q^{(i+1)}(\mu) = -\frac{1}{2} \langle \lambda \rangle_{q^{(i)}(\lambda)} \sum_{n=1}^N (y_n - \mu)^2 - \frac{1}{2} \frac{(\mu - m_\mu)^2}{s_\mu^2} + \tilde{C}_\mu \quad (5.27)$$

we first note that

$$\langle \lambda \rangle_{q^{(i)}(\lambda)} = a_\lambda^{(i)} b_\lambda^{(i)} \quad (5.28)$$

and that

$$\sum_{n=1}^N (y_n - \mu)^2 = \sum_{n=1}^N (y_n^2 - 2y_n\mu + \mu^2) = \sum_{n=1}^N y_n^2 - 2\mu \sum_{n=1}^N y_n + N\mu^2 \quad (5.29)$$

Substitution then yields

$$\begin{aligned} \ln q^{(i+1)}(\mu) &= -\frac{1}{2} \langle \lambda \rangle_{q^{(i)}(\lambda)} \sum_{n=1}^N (y_n - \mu)^2 - \frac{1}{2} \frac{(\mu - m_\mu)^2}{s_\mu^2} + \tilde{C}_\mu \\ &= -\frac{1}{2} a_\lambda^{(i)} b_\lambda^{(i)} (\sum_{n=1}^N y_n^2 - 2\mu \sum_{n=1}^N y_n + N\mu^2) - \frac{1}{2} \frac{(\mu - m_\mu)^2}{s_\mu^2} + \tilde{C}_\mu \\ &= -\frac{1}{2} a_\lambda^{(i)} b_\lambda^{(i)} (\sum_{n=1}^N y_n^2 - 2\mu \sum_{n=1}^N y_n + N\mu^2) - \frac{1}{2} \frac{\mu^2 - 2\mu m_\mu + m_\mu^2}{s_\mu^2} + \tilde{C}_\mu \\ &= -\frac{1}{2} \left(a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n^2 - a_\lambda^{(i)} b_\lambda^{(i)} 2\mu \sum_{n=1}^N y_n + a_\lambda^{(i)} b_\lambda^{(i)} N\mu^2 + \frac{1}{s_\mu^2} \mu^2 - \frac{2m_\mu}{s_\mu^2} \mu + \frac{1}{s_\mu^2} m_\mu^2 \right) + \tilde{C}_\mu \end{aligned} \quad (5.30)$$

Reordering the last expression on the right-hand side in terms of powers of μ , we obtain

$$\begin{aligned} \ln q^{(i+1)}(\mu) &= -\frac{1}{2} \left(a_\lambda^{(i)} b_\lambda^{(i)} N\mu^2 + \frac{1}{s_\mu^2} \mu^2 - a_\lambda^{(i)} b_\lambda^{(i)} 2\mu \sum_{n=1}^N y_n - \frac{2m_\mu}{s_\mu^2} \mu + a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n^2 + \frac{1}{s_\mu^2} m_\mu^2 \right) + \tilde{C}_\mu \\ &= -\frac{1}{2} \left(\left(a_\lambda^{(i)} b_\lambda^{(i)} N + \frac{1}{s_\mu^2} \right) \mu^2 - \left(2a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n + \frac{2m_\mu}{s_\mu^2} \right) \mu + a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n^2 + \frac{1}{s_\mu^2} m_\mu^2 \right) + \tilde{C}_\mu \end{aligned} \quad (5.31)$$

Resolving the brackets, and grouping terms independent of μ with the constant $\tilde{C}_\mu$, resulting in the new constant $\tilde{\tilde{C}}_\mu$, then simplifies the above to

$$\ln q^{(i+1)}(\mu) = -\frac{1}{2} \left(a_\lambda^{(i)} b_\lambda^{(i)} N + \frac{1}{s_\mu^2} \right) \mu^2 + \left(a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n + \frac{m_\mu}{s_\mu^2} \right) \mu + \tilde{\tilde{C}}_\mu \quad (5.32)$$

Finally, we may re-express the coefficient of μ^2 more conveniently by setting

$$a_\lambda^{(i)} b_\lambda^{(i)} N + \frac{1}{s_\mu^2} = \frac{a_\lambda^{(i)} b_\lambda^{(i)} N s_\mu^2}{s_\mu^2} + \frac{1}{s_\mu^2} = \frac{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}}{s_\mu^2} \quad (5.33)$$

and obtain

$$\ln q^{(i+1)}(\mu) = -\frac{1}{2} \left(\frac{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}}{s_\mu^2} \right) \mu^2 + \left(a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n + \frac{m_\mu}{s_\mu^2} \right) \mu + \tilde{C}_\mu \quad (5.34)$$

Using the completing-the-square theorem in the form

$$\exp \left(-\frac{1}{2} a x^2 + b x \right) = N(x; a^{-1} b, a^{-1}) \cdot C \quad (5.35)$$

then yields

$$q^{(i+1)}(\mu) \propto N \left(\mu; m_\mu^{(i+1)}, s_\mu^{2(i+1)} \right) \quad (5.36)$$

where the variational parameters $m_\mu^{(i+1)}, s_\mu^{2(i+1)}$ may be expressed in terms of the i th variational parameters of $q^{(i)}(\lambda)$, the prior parameters m_μ and s_μ^2 and the data y_n as

$$s_\mu^{2(i+1)} = \left(\frac{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}}{s_\mu^2} \right)^{-1} = \frac{s_\mu^2}{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}} \quad (5.37)$$

and

$$\begin{aligned} m_\mu^{(i+1)} &= \frac{s_\mu^2}{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}} \left(a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n + \frac{m_\mu}{s_\mu^2} \right) \\ &= \frac{s_\mu^2}{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}} \left(\frac{m_\mu + a_\lambda^{(i)} b_\lambda^{(i)} s_\mu^2 \sum_{n=1}^N y_n}{s_\mu^2} \right) \\ &= \frac{m_\mu + s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)} \sum_{n=1}^N y_n}{1 + N s_\mu^2 a_\lambda^{(i)} b_\lambda^{(i)}} \end{aligned} \quad (5.39)$$

□

5.4 Variational parameter update-equation derivation for $q(\lambda)$

The variational inference theorem for approximated posteriors (5.13) states that the logarithm of the optimal distribution over the unobservable variable λ , given a momentarily fixed distribution over the unobservable variable μ , is obtained by setting (with a constant $C_\lambda \in \mathbb{R}$)

$$\ln q^{(i+1)}(\lambda) := \langle \ln p(y, \mu, \lambda) \rangle_{q^{(i+1)}(\mu)} + \ln C_\lambda \quad (5.40)$$

which, if we identify the prior distribution $p(\mu)p(\lambda)$ with the initial variational distribution $q^{(0)}(x)q^{(0)}(\lambda)$, may be rewritten as

$$\ln q^{(i+1)}(\lambda) = \frac{N}{2} \ln \lambda - \frac{\lambda}{2} \langle \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i+1)}(\mu)} + (a_\lambda - 1) \ln \lambda - \frac{\lambda}{b_\lambda} + \tilde{C}_\lambda \quad (5.41)$$

Further, upon evaluation of the expectation and taking the exponential of (5.13), it follows that $q^{(i+1)}(\lambda)$ is proportional to a Gamma distribution

$$G \left(\lambda; a_\lambda^{(i+1)}, b_\lambda^{(i+1)} \right) \quad (5.42)$$

with parameters

$$a_\lambda^{(i+1)} := \frac{N}{2} + a_\lambda \text{ and } b_\lambda^{(i+1)} := \left(\frac{1}{b_\lambda} + \frac{1}{2} \left(\sum_{n=1}^N y_n^2 - 2 \sum_{n=1}^N y_n m_\mu^{(i+1)} + N \left((m_\mu^{(i+1)})^2 + s_\mu^{2(i+1)} \right) \right) \right)^{-1} \quad (5.43)$$

○ **Derivation of equation (5.40)**

In analogy to the derivation of equations (5.15) we have

$$\begin{aligned} \ln q^{(i+1)}(\lambda) &= \langle \ln \left(\prod_{n=1}^N \left(\frac{\lambda}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\lambda}{2} (y_n - \mu)^2 \right) \right) \rangle_{q^{(i+1)}(\mu)} \\ &\quad + \langle \ln \left(\left(2\pi s_\mu^{2(i+1)} \right)^{-\frac{1}{2}} \exp \left(-\frac{1}{2s_\mu^{2(i+1)}} \left(\mu - m_\mu^{(i+1)} \right)^2 \right) \right) \rangle_{q^{(i+1)}(\mu)} \\ &\quad + \langle \ln \left(\frac{1}{\Gamma(a_\lambda)} \frac{1}{(b_\lambda)^{a_\lambda}} \lambda^{a_\lambda-1} \exp \left(-\frac{\lambda}{b_\lambda} \right) \right) \rangle_{q^{(i+1)}(\mu)} \\ &\quad + \ln C_\lambda \\ &= \langle \frac{N}{2} \ln \lambda - \frac{1}{2} \ln 2\pi - \frac{\lambda}{2} \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i+1)}(\mu)} \\ &\quad + \langle -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln s_\mu^{2(i+1)} - \frac{(\mu - m_\mu^{(i+1)})^2}{2s_\mu^{2(i+1)}} \rangle_{q^{(i+1)}(\mu)} \\ &\quad + \langle -\ln(\Gamma(a_\lambda)) - \ln((b_\lambda)^{a_\lambda}) + (a_\lambda - 1) \ln \lambda - \frac{\lambda}{b_\lambda} \rangle_{q^{(i+1)}(\mu)} \\ &\quad + \ln C_\lambda \end{aligned} \quad (5.44)$$

Grouping all constant terms independent of λ (i.e. those terms that do not change if λ changes) in a constant $\tilde{C}_\lambda \in \mathbb{R}$ then simplifies the above to

$$\begin{aligned} \ln q^{(i+1)}(\lambda) &= \langle \frac{N}{2} \ln \lambda - \frac{\lambda}{2} \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i+1)}(\mu)} + \langle (a_\lambda - 1) \ln \lambda - \frac{\lambda}{b_\lambda} \rangle_{q^{(i+1)}(\mu)} + \tilde{C}_\lambda \\ &= \frac{N}{2} \ln \lambda - \frac{\lambda}{2} \langle \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i+1)}(\mu)} + (a_\lambda - 1) \ln \lambda - \frac{\lambda}{b_\lambda} + \tilde{C}_\lambda \end{aligned} \quad (5.45)$$

□

○ **Derivation of equations (5.41) and (5.42)**

Reorganizing the right hand side of equation (5.40) in multiplicative terms involving λ yields

$$\begin{aligned} \ln q^{(i+1)}(\lambda) &= \left(\frac{N}{2} + a_\lambda - 1 \right) \ln \lambda - \left(\frac{1}{b_\lambda} + \frac{1}{2} \langle \sum_{n=1}^N (y_n - \mu)^2 \rangle_{q^{(i+1)}(\mu)} \right) \lambda + \tilde{C}_\lambda \\ &= \left(\frac{N}{2} + a_\lambda - 1 \right) \ln \lambda - \left(\frac{1}{b_\lambda} + \frac{1}{2} \langle \sum_{n=1}^N y_n^2 - 2\mu \sum_{n=1}^N y_n + N\mu^2 \rangle_{q^{(i+1)}(\mu)} \right) \lambda + \tilde{C}_\lambda \\ &= \left(\frac{N}{2} + a_\lambda - 1 \right) \ln \lambda - \left(\frac{1}{b_\lambda} + \frac{1}{2} \left(\sum_{n=1}^N y_n^2 - 2 \sum_{n=1}^N y_n \langle \mu \rangle_{q^{(i+1)}(\mu)} + N \langle \mu^2 \rangle_{q^{(i+1)}(\mu)} \right) \right) \lambda + \tilde{C}_\lambda \end{aligned} \quad (5.46)$$

where the expectations may be expressed in terms of the variational parameters of $q^{(i+1)}(\mu)$ according to

$$\ln q^{(i+1)}(\lambda) = \left(\frac{N}{2} + a_\lambda - 1\right) \ln \lambda - \left(\frac{1}{b_\lambda} + \frac{1}{2} \left(\sum_{n=1}^N y_n^2 - 2 \sum_{n=1}^N y_n m_\mu^{(i+1)} + N \left((m_\mu^{(i+1)})^2 + s_\mu^{2(i+1)} \right) \right)\right) \lambda + \tilde{C}_\lambda \quad (5.47)$$

Taking the exponential on both sides then yields

$$q^{(i+1)}(\lambda) \propto \lambda^{\left(\frac{N}{2} + a_\lambda - 1\right)} \exp \left(\left(-\frac{1}{b_\lambda} - \frac{1}{2} \left(\sum_{n=1}^N y_n^2 - 2 \sum_{n=1}^N y_n m_\mu^{(i+1)} + N \left((m_\mu^{(i+1)})^2 + s_\mu^{2(i+1)} \right) \right) \right) \lambda \right) \quad (5.48)$$

Up to a normalization constant $\tilde{C}_\lambda$, $q^{(i+1)}(\lambda)$ is thus given by a Gamma distribution with

$$q^{(i+1)} \propto \lambda^{a_\lambda^{(i+1)} - 1} \exp \left(-\frac{\lambda}{b_\lambda^{(i+1)}} \right) \quad (5.49)$$

where

$$a_\lambda^{(i+1)} := \frac{N}{2} + a_\lambda \quad (5.50)$$

and

$$b_\lambda^{(i+1)} := \left(\frac{1}{b_\lambda} + \frac{1}{2} \left(\sum_{n=1}^N y_n^2 - 2 \sum_{n=1}^N y_n m_\mu^{(i+1)} + N \left((m_\mu^{(i+1)})^2 + s_\mu^{2(i+1)} \right) \right) \right)^{-1} \quad (5.51)$$

□

5.6 Evaluation of the variational free energy

While the maximization of the variational free energy has been exploited to derive the variational parameter update equations, the value of the variational free energy has not yet been directly evaluated. To obtain a value of this (implicit) target function to monitor the evolution of the algorithm, and, more importantly, to obtain an approximation to the marginal likelihood upon convergence, it is desirable to evaluate the variational free energy integral (cf. equation (13) of the tutorial). We first reformulate the variational free energy as follows (Penny, 2000)

$$\mathcal{F}(q(\vartheta)) = \int q(\vartheta) \ln(p(y|\vartheta)) d\vartheta - \int q(\vartheta) \ln \left(\frac{q(\vartheta)}{p(\vartheta)} \right) d\vartheta := \mathcal{L}_{av}(p(y|\vartheta), q(\vartheta)) - \mathcal{KL}(q(\vartheta)||p(\vartheta)) \quad (5.52)$$

The first term in the expression on the right hand side of (5.53) is sometimes referred to as the “average likelihood” and the second term is the KL-divergence between the variational and prior distributions.

◦ Derivation of (5.52)

By definition of the variational free energy, we have

$$\begin{aligned} \mathcal{F}(q(\vartheta)) &= \int q(\vartheta) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta)} \right) d\vartheta \\ &= \int q(\vartheta) \ln \left(\frac{p(y|\vartheta)p(\vartheta)}{q(\vartheta)} \right) d\vartheta \\ &= \int q(\vartheta) \ln(p(y|\vartheta)) - \ln \left(\frac{q(\vartheta)}{p(\vartheta)} \right) d\vartheta \\ &= \int q(\vartheta) \ln(p(y|\vartheta)) d\vartheta - \int q(\vartheta) \ln \left(\frac{q(\vartheta)}{p(\vartheta)} \right) d\vartheta \end{aligned}$$

$$=: \mathcal{L}_{av}(p(y|\vartheta), q(\vartheta)) - \mathcal{KL}(q(\vartheta)||p(\vartheta)) \quad (5.53)$$

□

Upon this reformulation the respective terms comprising the variational free energy may then be evaluated in turn as shown below, resulting in

$$\mathcal{F}\left(q^{(i)}(\mu)q^{(i)}(\lambda)\right) := \mathcal{L}_{av}\left(p(y|\vartheta), q^{(i)}(\mu)q^{(i)}(\lambda)\right) - \mathcal{KL}(q^{(i)}(\mu)q^{(i)}(\lambda)||p(\mu)p(\lambda)) \quad (5.54)$$

where

$$\mathcal{L}_{av}\left(p(y|\vartheta), q^{(i)}(\mu)q^{(i)}(\lambda)\right) := \frac{1}{2}(\psi(a_\lambda^{(i)}) + \ln(b_\lambda^{(i)})) - \frac{1}{2}a_\lambda^{(i)}b_\lambda^{(i)}\left(\sum_{n=1}^N y_n^2 + N\left((m_\mu^{(i)})^2 + s_\mu^{2(i)}\right) - 2m_\mu^{(i)}\sum_{n=1}^N y_n\right) \quad (5.55)$$

and

$$\mathcal{KL}(q^{(i)}(\mu)q^{(i)}(\lambda)||p(\mu)p(\lambda)) = \mathcal{KL}(q^{(i)}(\mu)||p(\mu)) + \mathcal{KL}(q^{(i)}(\lambda)||p(\lambda)) \quad (5.56)$$

with

$$\begin{aligned} \mathcal{KL}(q^{(i)}(\mu)||p(\mu)) &= \frac{1}{2}\ln\frac{s_\mu^2}{s_\mu^{2(i)}} + \frac{m_\mu^2 + m_\mu^{(i)2} + s_\mu^{2(i)} - 2m_\mu^{(i)}m_\mu}{2s_\mu^2} - \frac{1}{2} \\ \mathcal{KL}(q^{(i)}(\lambda)||p(\lambda)) &= (b_\lambda^{(i)} - 1)\psi(b_\lambda^{(i)}) - \ln a_\lambda^{(i)} - b_\lambda^{(i)} - \log \Gamma(b_\lambda^{(i)}) + \ln \Gamma(b_\lambda) + b_\lambda \ln a_\lambda - (b_\lambda - 1)(\psi(b_\lambda^{(i)}) + \ln a_\lambda^{(i)}) + \frac{a_\lambda^{(i)}b_\lambda^{(i)}}{a_\lambda} \end{aligned} \quad (5.57)$$

We thus obtain an expression for the variational free energy as function of the observable data $y_1, \dots, y_N$, the variational parameters $m_\mu^{(i)}, s_\mu^{2(i)}, a_\lambda^{(i)}, b_\lambda^{(i)}$ and the prior parameters $m_\mu, s_\mu^2, a_\lambda$ and b_λ .

○ Derivation of (5.54)

We first recall the definitions of $p(y|\vartheta), p(\vartheta)$ and $q(\vartheta)$ for the current example:

$$p(y|\vartheta) = p(y|\mu, \lambda) = \prod_{n=1}^N N(y_n; \mu, \lambda) \quad (5.58)$$

$$p(\vartheta) = p(\mu, \lambda) =: p(\mu)p(\lambda) = N(\mu; m_\mu, s_\mu^2)G(\lambda; a_\lambda, b_\lambda) \quad (5.59)$$

and

$$q(\vartheta) = q(\mu, \lambda) =: q(\mu)q(\lambda) = N(\mu; m_\mu^{(i)}, s_\mu^{2(i)})G(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)}) \quad (5.60)$$

(a) Evaluation of $\mathcal{L}_{av}\left(p(y|\vartheta), q^{(i)}(\mu)q^{(i)}(\lambda)\right)$ in (5.55)

Substitution of (5.58) and (5.59) in (5.54) yields,

$$\begin{aligned} \int q(\vartheta) \ln(p(y|\vartheta)) d\vartheta &= \iint q^{(i)}(\mu)q^{(i)}(\lambda) \ln\left(\prod_{n=1}^N N(y_n; \mu, \lambda)\right) d\mu d\lambda \\ &= \iint q^{(i)}(\mu)q^{(i)}(\lambda) \ln\left(\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \prod_{n=1}^N \exp\left(-\frac{\lambda}{2}(y_n - \mu)^2\right)\right) d\mu d\lambda \end{aligned}$$

$$\begin{aligned}
&= \iint q^{(i)}(\mu)q^{(i)}(\lambda) \left(\frac{1}{2} \ln \lambda - \frac{1}{2} \ln 2\pi - \frac{\lambda}{2} \sum_{n=1}^N (y_n - \mu)^2 \right) d\mu d\lambda \\
&= \frac{1}{2} \iint q^{(i)}(\mu)q^{(i)}(\lambda) (\ln \lambda) d\mu d\lambda - \iint q^{(i)}(\mu)q^{(i)}(\lambda) \left(\frac{\lambda}{2} \sum_{n=1}^N (y_n - \mu)^2 \right) d\mu d\lambda - \frac{1}{2} \ln 2\pi \\
&= \frac{1}{2} \int q^{(i)}(\lambda) \ln \lambda d\lambda - \int q^{(i)}(\lambda) \frac{\lambda}{2} \left(\int q^{(i)}(\mu) (\sum_{n=1}^N (y_n - \mu)^2) d\mu \right) d\lambda - \frac{1}{2} \ln 2\pi
\end{aligned} \tag{5.61}$$

The first integral term above is the expectation of a logarithm variable λ under the variational Gamma distribution $G\left(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)}\right)$ which can be shown to evaluate to

$$\int q^{(i)}(\lambda) \ln \lambda d\lambda = \psi\left(a_\lambda^{(i)}\right) + \ln b_\lambda^{(i)} \tag{5.62}$$

where

$$\psi: \mathbb{R} \rightarrow \mathbb{R}, x \mapsto \psi(x) := \frac{d}{dx} \Gamma(x) \tag{5.63}$$

denotes the di-gamma function (see also 2.34). The second integral term in (5.61) evaluates to

$$\begin{aligned}
\int q^{(i)}(\lambda) \frac{1}{2} \lambda \left(\int q^{(i)}(\mu) (\sum_{n=1}^N (y_n - \mu)^2) d\mu \right) d\lambda &= \int q^{(i)}(\lambda) \left(\frac{1}{2} \lambda \right) \left(\int q^{(i)}(\mu) (\sum_{n=1}^N y_n^2 - 2\mu \sum_{n=1}^N y_n + N\mu^2) d\mu \right) d\lambda \\
&= \frac{1}{2} \int q^{(i)}(\lambda) \lambda \left(\sum_{n=1}^N y_n^2 - 2 \sum_{n=1}^N y_n \int q^{(i)}(\mu) \mu d\mu + N \int q^{(i)}(\mu) \mu^2 d\mu \right) d\lambda \\
&= \frac{1}{2} a_\lambda^{(i)} b_\lambda^{(i)} \left(\sum_{n=1}^N y_n^2 - 2 m_\mu^{(i)} \sum_{n=1}^N y_n + N \left(\left(m_\mu^{(i)} \right)^2 + s_\mu^{2(i)} \right) \right)
\end{aligned} \tag{5.64}$$

(b) Evaluation of $\mathcal{KL}(q(\vartheta)||p(\vartheta))$ in (5.54)

We first note that we have, omitting superscripts for ease of notation,

$$\begin{aligned}
\mathcal{KL}(q(\mu)q(\lambda)||p(\mu)p(\lambda)) &= \int q(\mu)q(\lambda) \ln \left(\frac{p(\mu)p(\lambda)}{q(\mu)q(\lambda)} \right) d\mu d\lambda \\
&= \int q(\mu)q(\lambda) \left(\ln \frac{p(\mu)}{q(\mu)} + \ln \frac{p(\lambda)}{q(\lambda)} \right) d\mu d\lambda \\
&= \int q(\mu) \left(\ln \frac{p(\mu)}{q(\mu)} \right) d\mu + \int q(\lambda) \left(\ln \frac{p(\lambda)}{q(\lambda)} \right) d\lambda \\
&= \mathcal{KL}(q(\mu)||p(\mu)) + \mathcal{KL}(q(\lambda)||p(\lambda))
\end{aligned} \tag{5.65}$$

In the current example, the variable μ is governed by Gaussian distribution both for the prior and in the variational approximation to the posterior distribution. The variational distribution $q(\mu)$ corresponds to the i th variational

Gaussian distribution over μ and has parameters $m_\mu^{(i)}, s_\mu^{2(i)}$ whereas the Gaussian prior distribution over μ has parameters m_μ, s_μ^2 . We thus have

$$\mathcal{KL}(q(\mu)||p(\mu)) = \mathcal{KL}(q^{(i)}(\mu)||p(\mu)) = \mathcal{KL}\left(N(\mu; m_\mu^{(i)}, s_\mu^{2(i)})||N(\mu; m_\mu, s_\mu^2)\right) \quad (5.66)$$

As noted in (Penny, 2000), the KL divergence for two Gaussian distributions is a function of the respective distribution parameters and given for the current example as

$$\mathcal{KL}\left(N(\mu; m_\mu^{(i)}, s_\mu^{2(i)})||N(\mu; m_\mu, s_\mu^2)\right) = \frac{1}{2} \ln \frac{s_\mu^2}{s_\mu^{2(i)}} + \frac{m_\mu^2 + m_\mu^{(i)2} + s_\mu^{2(i)} - 2m_\mu^{(i)}m_\mu}{2s_\mu^2} - \frac{1}{2} \quad (5.67)$$

Finally, for the current example, the variable λ is governed by a Gamma distribution both for the prior and in the variational approximation to the posterior distribution. The variational distribution $q(\lambda)$ corresponds to the i th variational Gamma distribution over λ and has parameters $a_\lambda^{(i)}, b_\lambda^{(i)}$, whereas the Gamma prior distribution over λ has parameters a_λ, b_λ . Hence, we have

$$\mathcal{KL}(q(\lambda)||p(\lambda)) = \mathcal{KL}(q^{(i)}(\lambda)||p(\lambda)) = \mathcal{KL}\left(G(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)})||G(\lambda; a_\lambda, b_\lambda)\right) \quad (5.68)$$

As noted in (Penny, 2000), the KL divergence for two Gamma distributions is a function of the respective Gamma distribution parameters and given for the current example as

$$\mathcal{KL}(G(\lambda; a_\lambda^{(i)}, b_\lambda^{(i)})||G(\lambda; a_\lambda, b_\lambda)) = (b_\lambda^{(i)} - 1)\psi(b_\lambda^{(i)}) - \ln a_\lambda^{(i)} - b_\lambda^{(i)} - \log \Gamma(b_\lambda^{(i)}) + \ln \Gamma(b_\lambda) + b_\lambda \ln a_\lambda - (b_\lambda - 1)(\psi(b_\lambda) + \ln a_\lambda) + \frac{a_\lambda^{(i)} b_\lambda^{(i)}}{a_\lambda} \quad (5.69)$$

Concatenating the results from (a) and (b) then results in expressions (5.55) for the variational free energy.

□

The VB algorithm derived in this example is shown in pseudocode form in Table 3.

```

% Initialization of the variational parameters and variational free energy
 $m_{\mu}^{(0)} := m_{\mu}$ 
 $s_{\mu}^{2(0)} := s_{\mu}^2$ 
 $a_{\lambda}^{(0)} := a_{\lambda}$ 
 $b_{\lambda}^{(0)} := b_{\lambda}$ 

 $F^{(0)} := \mathcal{F}\left(y, N, m_{\mu}^{(0)}, s_{\mu}^{2(0)}, a_{\lambda}^{(0)}, b_{\lambda}^{(0)}, m_{\mu}, s_{\mu}^2, a_{\lambda}, b_{\lambda}\right)$ 

% VB iterations
for  $i = 0, 1, 2, \dots$  until convergence do

    % update of  $q(\mu)$ 
 $m_{\mu}^{(i+1)} := \frac{m_{\mu} + s_{\mu}^2 a_{\lambda}^{(i)} b_{\lambda}^{(i)} \sum_{n=1}^N y_n}{1 + N s_{\mu}^2 a_{\lambda}^{(i)} b_{\lambda}^{(i)}}$ 
 $s_{\mu}^{2(i+1)} := \frac{s_{\mu}^2}{1 + N s_{\mu}^2 a_{\lambda}^{(i)} b_{\lambda}^{(i)}}$ 

    % update of  $q(\lambda)$ 
 $a_{\lambda}^{(i+1)} := \frac{N}{2} + a_{\lambda}$ 
 $b_{\lambda}^{(i+1)} := \left( \frac{1}{b_{\lambda}} + \frac{1}{2} \left( \sum_{n=1}^N y_n^2 - 2 \sum_{n=1}^N y_n m_{\mu}^{(i+1)} + N \left( \left( m_{\mu}^{(i+1)} \right)^2 + s_{\mu}^{2(i+1)} \right) \right) \right)^{-1}$ 

    % Free Energy Evaluation
 $F^{(i+1)} := \mathcal{F}\left(y, N, m_{\mu}^{(0)}, s_{\mu}^{2(0)}, a_{\lambda}^{(0)}, b_{\lambda}^{(0)}, m_{\mu}, s_{\mu}^2, a_{\lambda}, b_{\lambda}\right)$ 

end

% VB posterior unobservable variable distribution and log model evidence approximation
 $p(\mu, \lambda | y)_{VB} := N\left(\mu; m_{\mu}^{(i+1)}, s_{\mu}^{2(i+1)}\right) \cdot G\left(\lambda; a_{\lambda}^{(i+1)}, b_{\lambda}^{(i+1)}\right)$ 
 $\ln p(y)_{VB} := F^{(i+1)}$ 

```

Table 3. Pseudocode for the VB Algorithm for the univariate Gaussian.

5.7 Matlab implementation

The following Matlab code implementing the pseudocode of Table 3 was used to generate Figure 5 of the tutorial.

```
function vb_lgssm_1

% This function demonstrates the variational Bayesian algorithm for the
% example of inferring the mean and precision (inverse variance) of a
% univariate Gaussian distribution as discussed in the tutorial section
% 3.2 and supplementary section 5.
%
% Copyright (C) Dirk Ostwald
% -----

clc
close all

% -----
%                               Random number generator settings
% -----
rng('shuffle')

% -----
%                               Parameter Space Initialization
% -----

% specify observed variable y space
ymax = 4 ;
ymin = -4 ;
yres = 100;
yspace = linspace(ymin, ymax, yres);

% specify unobserved variable mu space
mumax = 2;
mumin = -2;
mures = 100;
muspace = linspace(mumin, mumax, mures);

% specify parameter variable lambda space
lmax = 7 ;
lmin = 0.001;
lres = 100 ;
lspace = linspace(lmax, lmin, lres);

% -----
%                               Specification of the true, but unknown, distribution
% -----
% true, but unknown, value of the latent variable x
mu_tbu = 1;

% true, but unknown, value of the parameter variable lambda
l_tbu = 5;

% true, but unknown, underlying model
y_tbu = ((l_tbu)/(2*pi))^(1/2) .* exp((-l_tbu/2) * ((yspace-mu_tbu).^2));

% -----
%                               IID sampling from the true, but unknown, model
% -----
% number of observed variable data points
N = 10;

% i.i.d. sampling
yn = NaN(N,1);
for i = 1:N
    yn(i) = mu_tbu + sqrt(1/l_tbu)*randn;
end

% -----
%                               Variational Bayes algorithm initialization
% -----
% generative model prior distribution
m_mu_prior = 0;
s_mu_prior = 1;
a_l_prior = 1;
b_l_prior = 1;

% maximum number of iterations
```

```

maxIter      = 8;

% initialization of parameter, variational distribution and FE arrays
m_mu        = NaN(maxIter+1,1);      % expectation of Gaussian over x
s_mu        = NaN(maxIter+1,1);      % variance of Gaussian over x
a_l         = NaN(maxIter+1,1);      % shape of Gamma over l
b_l         = NaN(maxIter+1,1);      % scale of Gamma over l
vFE_it      = NaN(maxIter+1,1);      % variational free energy
qmuql       = NaN(mures,lres,maxIter+1); % q(mu)q(l) on muspace x lspace support

% initialization of the variational distribution parameters
m_mu(1)     = m_mu_prior;
s_mu(1)     = s_mu_prior;
a_l(1)      = a_l_prior;
b_l(1)      = b_l_prior;

% evaluate the factorized Gaussian-Gamma distribution over x and lambda
qmuql(:, :, 1) = pdf_qmuql(muspace, lspace, m_mu(1), s_mu(1), a_l(1), b_l(1));

% evaluate the expectations of the marginal distributions and the
% variational estimate of the true, but unknown, model based on these
% expectations
exp_mu_qmu(1) = m_mu(1);
exp_l_ql(1)   = a_l(1)*b_l(1);
exp_y_qmul(1,:) = ((exp_l_ql)/(2*pi))^(1/2)).*exp((-exp_l_ql/2)*((yspace-exp_mu_qmu).^2));

% data statistics
y_bar        = sum(yn,1);
yy_bar       = sum(yn.^2,1);

% evaluate the variational free energy upon initialization
vFE_it(1) = vFE(N,y_bar, yy_bar, m_mu(1), s_mu(1), m_mu(1), s_mu(1), a_l(1), b_l(1), a_l(1), b_l(1));

% -----
%                               Variational Bayesian algorithm iterations
% -----
for i = 1:maxIter

    % q(\mu) update
    % -----
    m_mu(i+1) = (m_mu_prior + (s_mu_prior*a_l(i)*b_l(i)*y_bar))/(1 + (N*s_mu_prior*a_l(i)*b_l(i)));
    s_mu(i+1) = s_mu_prior/(1 + (N*s_mu_prior*a_l(i)*b_l(i)));

    % evaluate the factorized Gaussian-Gamma after update of q(x)
    qmuql(:, :, 2*i) = pdf_qmuql(muspace, lspace, m_mu(i+1), s_mu(i+1), a_l(i), b_l(i));

    % evaluate the expectations of the marginal distributions and the
    % variational estimate of the true, but unknown, model based on these
    % expectations
    exp_mu_qmu = m_mu(i+1);
    exp_l_ql   = a_l(i)*b_l(i);
    exp_y_qmul(2*i,:) = ((exp_l_ql)/(2*pi))^(1/2)).*exp((-exp_l_ql/2)*((yspace-exp_mu_qmu).^2));

    % q(\lambda) update
    % -----
    a_l(i+1) = (N/2) + a_l_prior;
    b_l(i+1) = 1/(1/b_l_prior + ((1/2)*(yy_bar - (2*m_mu(i+1)*y_bar) + (N*(m_mu(i+1)^2 + s_mu(i+1))))));

    % evaluate the factorized Gaussian-Gamma after update of q(x)
    qmuql(:, :, 2*i+1) = pdf_qmuql(muspace, lspace, m_mu(i+1), s_mu(i+1), a_l(i+1), b_l(i+1));

    % evaluate the expectations of the marginal distributions and the
    % variational estimate of the true, but unknown, model based on these
    % expectations
    exp_mu_qmu = m_mu(i+1);
    exp_l_ql   = a_l(i+1)*b_l(i+1);
    exp_y_qmul(2*i+1,:) = ((exp_l_ql)/(2*pi))^(1/2)).*exp((-exp_l_ql/2)*((yspace-exp_mu_qmu).^2));

    % evaluation of the variational free energy
    % -----
    vFE_it(i+1) = vFE(N,y_bar, yy_bar, m_mu(i+1), s_mu(i+1), m_mu_prior, s_mu_prior, a_l(i+1),
    b_l(i+1), a_l_prior, b_l_prior);

end

% -----
%                               Result plotting
% -----
h = figure;
set(h, 'Color', [1 1 1])

% initialization

```

```

subplot(3,5,1)
hold on
plot(yspace, y_tbu, 'k-', 'LineWidth', 2)
plot(yspace, exp_y_qmul(1,:), 'k:', 'LineWidth', 2)
plot(yn, zeros(N,1), 'ro', 'MarkerSize', 6, 'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r')
axis square
set(gca, 'FontSize', 16, 'LineWidth', 2, 'FontName', 'Times New Roman')

subplot(3,5,6)
hold on
contourf(muspace, lspace, qmuql(:, :, 1))
set(gca, 'FontSize', 16, 'LineWidth', 2, 'YDir', 'Normal', 'FontName', 'Times New Roman')
plot(mu_tbu, l_tbu, 'wo', 'MarkerSize', 6, 'MarkerFaceColor', 'w')
axis square

% first iteration
subplot(3,5,2)
hold on
plot(yspace, y_tbu, 'k-', 'LineWidth', 2)
plot(yspace, exp_y_qmul(2,:), 'k:', 'LineWidth', 2)
plot(yn, zeros(N,1), 'ro', 'MarkerSize', 6, 'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r')
axis square
set(gca, 'FontSize', 16, 'LineWidth', 2, 'FontName', 'Times New Roman')

subplot(3,5,7)
hold on
contourf(muspace, lspace, qmuql(:, :, 2))
set(gca, 'FontSize', 16, 'LineWidth', 2, 'YDir', 'Normal', 'FontName', 'Times New Roman')
plot(mu_tbu, l_tbu, 'wo', 'MarkerSize', 6, 'MarkerFaceColor', 'w')
axis square

subplot(3,5,3)
hold on
plot(yspace, y_tbu, 'k-', 'LineWidth', 2)
plot(yspace, exp_y_qmul(3,:), 'k:', 'LineWidth', 2)
plot(yn, zeros(N,1), 'ro', 'MarkerSize', 6, 'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r')
axis square
set(gca, 'FontSize', 16, 'LineWidth', 2, 'FontName', 'Times New Roman')

subplot(3,5,8)
hold on
contourf(muspace, lspace, qmuql(:, :, 3))
set(gca, 'FontSize', 16, 'LineWidth', 2, 'YDir', 'Normal', 'FontName', 'Times New Roman')
plot(mu_tbu, l_tbu, 'wo', 'MarkerSize', 6, 'MarkerFaceColor', 'w')
axis square

% last iteration
subplot(3,5,4)
hold on
plot(yspace, y_tbu, 'k-', 'LineWidth', 2)
plot(yspace, exp_y_qmul(end-1,:), 'k:', 'LineWidth', 2)
plot(yn, zeros(N,1), 'ro', 'MarkerSize', 6, 'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r')
axis square
set(gca, 'FontSize', 16, 'LineWidth', 2, 'FontName', 'Times New Roman')

subplot(3,5,9)
hold on
contourf(muspace, lspace, qmuql(:, :, end-1))
set(gca, 'FontSize', 16, 'LineWidth', 2, 'YDir', 'Normal', 'FontName', 'Times New Roman')
plot(mu_tbu, l_tbu, 'wo', 'MarkerSize', 6, 'MarkerFaceColor', 'w')
axis square

subplot(3,5,5)
hold on
plot(yspace, y_tbu, 'k-', 'LineWidth', 2)
plot(yspace, exp_y_qmul(end,:), 'k:', 'LineWidth', 2)
plot(yn, zeros(N,1), 'ro', 'MarkerSize', 6, 'MarkerFaceColor', 'r', 'MarkerEdgeColor', 'r')
axis square
set(gca, 'FontSize', 16, 'LineWidth', 2, 'FontName', 'Times New Roman')

subplot(3,5,10)
hold on
contourf(muspace, lspace, qmuql(:, :, end))
set(gca, 'FontSize', 16, 'LineWidth', 2, 'YDir', 'Normal', 'FontName', 'Times New Roman')
plot(mu_tbu, l_tbu, 'wo', 'MarkerSize', 6, 'MarkerFaceColor', 'w')
axis square

maximize(h)
subplot(3,5,[11:15])
plot([0:maxIter], vFE_it(1:end), 'ko-', 'MarkerFaceColor', 'k', 'MarkerSize', 8, 'LineWidth', 2)
set(gca, 'FontSize', 24, 'LineWidth', 2, 'FontName', 'Times New Roman')
xlim([0 maxIter])

```

```

box off

end

function [qmuql] = pdf_qmuql(muspace, lspace, m_mu, s_mu, a_l, b_l)

% This function evaluates the factorized variational probability density
% function over a Gaussian distributed variable mu and a Gamma distributed
% variable lambda on the support muspace x lspace according to
%
%           p(mu,l; m_mu, s_mu, a_l, b_l) = N(mu;m_mu,s_mu)*G(l;a_l,b_l)
%
%
% Inputs
%       muspace      : support of variable mu (μ)
%       lspace       : support of variable l (lambda)
%       m_mu         : scalar expectation parameter of N(mu;m_mu,s_mu)
%       s_mu         : scalar variance parameter of N(mu;m_mu,s_mu)
%       a_l          : scalar shape parameter of G(l;a_l,b_l)
%       b_l          : scalar scale parameter of G(l;a_l,b_l)
%
% Outputs
%       qmuql        : p(mu,l; m_mu,s_mu,a_l, b_l)
%
% Copyright (C) Dirk Ostwald
% -----
% evaluate the factorized distribution q(mu)q(l)
qmuql = NaN(length(muspace), length(lspace));
for x = 1:length(muspace)
    for l = 1:length(lspace)
        qmuql(x,l) = (1/(sqrt(2*pi*s_mu)))*exp((-1/(2*s_mu))*(muspace(x) - m_mu)^2)*...
            ((1/gamma(a_l))*(lspace(l)^(a_l-1))/(b_l^a_l))*exp(-lspace(l)/b_l);
    end
end
end

function [vFE] = vFE(N,y_bar, yy_bar, m_mu, s_mu, m_mu_0, s_mu_0, a_l, b_l,a_l_0, b_l_0)

% This function evaluates the variational free energy as an approximation
% to the log marginal likelihood (log model evidence) of the example as
% discussed in supplement section 6.
%
% Inputs
%       N           : number of data points
%       y_bar       : sum of data points y_n
%       yy_bar      : sum of squared data points y_n^2
%       m_mu        : ith expectation parameter of unobserved variable mu variational distribution
%       s_mu        : ith variance parameter of unobserved variable mu variational distribution
%       m_mu_0      : expectation parameter initialization of unobserved variable mu variational distribution
%       s_mu_0      : expectation parameter initialization of unobserved variable mu variational distribution
%       a_l         : ith shape parameter of parameter variable \lambda variational distribution
%       b_l         : ith scale parameter of parameter variable \lambda variational distribution
%       a_l_0       : shape parameter initialization of parameter variable lambda variational distribution
%       b_l_0       : scale parameter initialization of parameter variable lambda variational distribution
%
% Outputs
%       vFE         : variational free energy
%
% Copyright (C) Dirk Ostwald
% -----
% termwise evaluation of the average energy
l_av = (1/2)*(psi(a_l) + log(b_l)) - ((1/2)*a_l*b_l*(yy_bar + N*(m_mu^2 + s_mu)) - (2*m_mu*y_bar));

% evaluation of the latent variable KL divergence
KL_q_mu = spm_kl_normal (m_mu,s_mu,m_mu_0,s_mu_0);

% evaluation of the parameter KL divergence
KL_q_l = spm_kl_gamma(b_l,a_l,b_l_0,a_l_0);

% concatenate the negative variational free energy
vFE = -(l_av - KL_q_mu - KL_q_l);

end

function [d] = spm_kl_gamma (b_q,c_q,b_p,c_p)
% KL divergence between two Gamma densities
% FORMAT [d] = spm_kl_gamma (b_q,c_q,b_p,c_p)
%
% KL (Q||P) = <log Q/P> where avg is wrt Q
%
% b_q, c_q      Parameters of first Gamma density
% b_p, c_p      Parameters of second Gamma density

```

```

%
% Copyright (C) 2008 Wellcome Trust Centre for Neuroimaging

% Will Penny
% $Id: spm_kl_gamma.m 2696 2009-02-05 20:29:48Z guillaume $

digamma_c_q = psi(c_q);
d           = (c_q-1)*digamma_c_q-log(b_q)-c_q-gamma1n(c_q);
d           = d+gamma1n(c_p)+c_p*log(b_p)-(c_p-1)*(digamma_c_q+log(b_q));
d           = d+b_q*c_q/b_p;

end

function [d] = spm_kl_normal (m_q,c_q,m_p,c_p)
% KL divergence between two multivariate normal densities
% FORMAT [d] = spm_kl_normal (m_q,c_q,m_p,c_p)
%
% KL (Q||P) = <log Q/P> where avg is wrt Q
%
% between two Normal densities Q and P
%
% m_q, c_q    Mean and covariance of first Normal density
% m_p, c_p    Mean and covariance of second Normal density
%
% Copyright (C) 2008 Wellcome Trust Centre for Neuroimaging

% Will Penny
% $Id: spm_kl_normal.m 2696 2009-02-05 20:29:48Z guillaume $

d=length(m_q);
m_q=m_q(:);
m_p=m_p(:);

Term1=0.5*spm_logdet(c_p)-0.5*spm_logdet(c_q);

inv_c_p=inv(c_p);
Term2=0.5*trace(inv_c_p*c_q)+0.5*(m_q-m_p)'*inv_c_p*(m_q-m_p);

d=Term1+Term2-0.5*d;

end

function [H] = spm_logdet(C)
% returns the log of the determinant of positive semi-definite matrix C
% FORMAT [H] = spm_logdet(C)
% H = log(det(C))
%
% spm_logdet is a computationally efficient operator that can deal with
% sparse matrices
%
% Copyright (C) 2008 Wellcome Trust Centre for Neuroimaging

% Karl Friston
% $Id: spm_logdet.m 4068 2010-09-07 16:50:08Z ged $

sw = warning('off','MATLAB:log:logOfZero');

% assume diagonal form
%-----
TOL = 1e-16; % c.f. n*max(s)*eps
n = length(C);
s = diag(C);
i = find(s > TOL & s < 1/TOL);
C = C(i,i);
H = sum(log(diag(C)));

% invoke det if non-diagonal
%-----
[i j] = find(C);
if any(i ~= j)
    n = length(C);
    a = exp(H/n);
    H = H + log(det(C/a));
end

% invoke svd if rank deficient
%-----
if ~isreal(H) || isinf(H)
    s = svd(full(C));
    H = sum(log(s(s > TOL & s < 1/TOL)));
end
warning(sw)

```

```
end
```

```
function maximize(fig)
```

```
% This function maximizes the figure with handle fig
```

```
%
```

```
% Input
```

```
%           fig      : figure handle of figure to be maximized
```

```
%
```

```
% Original author: Bill Finger, Create Inc.
```

```
% Free for redistribution as long as credit comments are preserved.
```

```
%
```

```
% Copyright (C) Dirk Ostwald
```

```
% -----
```

```
if nargin == 0
```

```
    fig = gcf;
```

```
end
```

```
units = get(fig,'units');
```

```
set(fig,'units','normalized','outerposition',[0 0 1 1]);
```

```
set(fig,'units',units);
```

```
end
```

6 Mathematical details of the tutorial example

6.1 VB for the univariate LGSSM

As outlined in the tutorial, we assume the following generative model

$$p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) = p(y_{1:T}, x_{1:T} | a, \lambda_x, b, \lambda_y) p(a, \lambda_x, b, \lambda_y) \quad (6.1)$$

where we define the following (conditional) independence properties

$$p(a, \lambda_x, b, \lambda_y) := p(a) p(\lambda_x) p(b) p(\lambda_y) \quad (6.2)$$

and

$$p(y_{1:T}, x_{1:T} | a, \lambda_x, b, \lambda_y) := p(x_1) \prod_{t=2}^T p(x_t | x_{t-1}, a, \lambda_x) \prod_{t=1}^T p(y_t | x_t, b, \lambda_y) \quad (6.3)$$

The marginal or prior probability density functions on the right-hand side of (6.2) are assumed to be given as

$$p(a) := N(a; \mu_a, \sigma_a^2) = (2\pi\sigma_a^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_a^2}(a - \mu_a)^2\right) \quad (6.4)$$

$$p(\lambda_x) := G(\lambda_x; a_{\lambda_x}, b_{\lambda_x}) = \frac{1}{\Gamma(a_{\lambda_x})} \frac{1}{b_{\lambda_x}^{a_{\lambda_x}}} \lambda_x^{a_{\lambda_x}-1} \exp\left(-\frac{\lambda_x}{b_{\lambda_x}}\right) \quad (6.5)$$

$$p(b) := N(b; \mu_b, \sigma_b^2) = (2\pi\sigma_b^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_b^2}(b - \mu_b)^2\right) \quad (6.6)$$

$$p(\lambda_y) := G(\lambda_y; a_{\lambda_y}, b_{\lambda_y}) = \frac{1}{\Gamma(a_{\lambda_y})} \frac{1}{b_{\lambda_y}^{a_{\lambda_y}}} \lambda_y^{a_{\lambda_y}-1} \exp\left(-\frac{\lambda_y}{b_{\lambda_y}}\right) \quad (6.7)$$

The marginal or prior distribution and the conditional distributions on the right-hand side of (6.3) are assumed to be given by

$$p(x_1) := N(x_1; \mu_{x_1}, \lambda_{x_1}^{-1}) = \left(\frac{\lambda_{x_1}}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda_{x_1}}{2}(x_1 - \mu_{x_1})^2\right) \quad (6.8)$$

$$p(x_t | x_{t-1}, a, \lambda_x) := N(x_t; ax_{t-1}, \lambda_x^{-1}) = \left(\frac{\lambda_x}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda_x}{2}(x_t - ax_{t-1})^2\right) \text{ for } t = 2, \dots, T \quad (6.9)$$

$$p(y_t | x_t, b, \lambda_y) := N(y_t; bx_t, \lambda_y^{-1}) = \left(\frac{\lambda_y}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda_y}{2}(y_t - bx_t)^2\right) \text{ for } t = 1, \dots, T \quad (6.10)$$

The set of known prior parameters is thus given by

$$\pi := \{\mu_a, \sigma_a^2, a_{\lambda_x}, b_{\lambda_x}, \mu_b, \sigma_b^2, a_{\lambda_y}, b_{\lambda_y}, \mu_{x_1}, \lambda_{x_1}\} \quad (6.11)$$

We assume the following factorized form of the variational approximation to the posterior distribution over the unobserved variables

$$p(a, \lambda_x, b, \lambda_y, x_{1:T} | y_{1:T}) \approx q(a) q(\lambda_x) q(b) q(\lambda_y) q(x_{1:T}) \quad (6.12)$$

According to the variational Bayes theorem for factorized posteriors (equation (17) of the tutorial), the variational free energy is maximized by setting

$$q(\vartheta_i) \propto \exp\left(\int q(\vartheta_{\setminus i}) \ln p(y, \vartheta) d\vartheta_{\setminus i}\right) \quad (6.13)$$

For the current example, we have

$$y := y_{1:T} \text{ and } \vartheta := \{a, \lambda_x, b, \lambda_y, x_{1:T}\} \quad (6.14)$$

where the variational distribution over unobserved variables partitions according to

$$q(\vartheta) := q(a)q(\lambda_x)q(b)q(\lambda_y)q(x_{1:T}) \quad (6.15)$$

We thus obtain the following update equations, using the short hand notation

$$\langle f(x) \rangle_{p(x)} := \int f(x)p(x)dx \quad (6.16)$$

for expectations:

$$q^{(i+1)}(a) \propto \exp \left(\langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T})} \right) \quad (6.17)$$

$$q^{(i+1)}(\lambda_x) \propto \exp \left(\langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(a)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T})} \right) \quad (6.18)$$

$$q^{(i+1)}(b) \propto \exp \left(\langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(a)q^{(i+1)}(\lambda_x)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T})} \right) \quad (6.19)$$

$$q^{(i+1)}(\lambda_y) \propto \exp \left(\langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(a)q^{(i+1)}(\lambda_x)q^{(i+1)}(b)q^{(i)}(x_{1:T})} \right) \quad (6.20)$$

$$q^{(i+1)}(x_{1:T}) \propto \exp \left(\langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(a)q^{(i+1)}(\lambda_x)q^{(i+1)}(b)q^{(i+1)}(\lambda_y)} \right) \quad (6.21)$$

For (6.17) – (6.20) we define the following variational distributions for VB iterations $i = 0, 1, 2, \dots$

$$q^{(i)}(a) := N(a; \mu_a^{(i)}, \sigma_a^{2(i)}) = \left(2\pi\sigma_a^{2(i)} \right)^{-\frac{1}{2}} \exp \left(-\frac{1}{2\sigma_a^{2(i)}} (a - \mu_a^{(i)})^2 \right) \quad (6.22)$$

$$q^{(i)}(\lambda_x) := G(\lambda_x; a_{\lambda_x}^{(i)}, b_{\lambda_x}^{(i)}) = \frac{1}{\Gamma(a_{\lambda_x}^{(i)})} \frac{1}{b_{\lambda_x}^{(i)} a_{\lambda_x}^{(i)}} \lambda_x^{a_{\lambda_x}^{(i)} - 1} \exp \left(-\frac{\lambda_x}{b_{\lambda_x}^{(i)}} \right) \quad (6.23)$$

$$q^{(i)}(b) := N(b; \mu_b^{(i)}, \sigma_b^{2(i)}) = \left(2\pi\sigma_b^{2(i)} \right)^{-\frac{1}{2}} \exp \left(-\frac{1}{2\sigma_b^{2(i)}} (b - \mu_b^{(i)})^2 \right) \quad (6.24)$$

$$q^{(i)}(\lambda_y) := G(\lambda_y; a_{\lambda_y}^{(i)}, b_{\lambda_y}^{(i)}) = \frac{1}{\Gamma(a_{\lambda_y}^{(i)})} \frac{1}{b_{\lambda_y}^{(i)} a_{\lambda_y}^{(i)}} \lambda_y^{a_{\lambda_y}^{(i)} - 1} \exp \left(-\frac{\lambda_y}{b_{\lambda_y}^{(i)}} \right) \quad (6.25)$$

Inference of $q^{(i+1)}(x_{1:T})$ is discussed in detail in Section 6.3 of the supplementary material. Note that the (known) prior parameters of (6.11) and the (to be evaluated) parameters of the variational distributions over unobserved variables are distinguished by the (i) superscript

6.2 Evaluation of $q(a)q(\lambda_x)q(b)q(\lambda_y)$

To derive the update equations for the parameters governing the variational distributions of the unobserved variables a, λ_x, b and λ_y , we take the following approach: We first reformulate the expectation of the joint distribution of the LGSSM and parameters, $p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y)$, under the variational mean field approximation $q^{(j)}(a)q^{(k)}(\lambda_x)q^{(l)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_{1:T})$. We can then in turn derive the update equation for each variational distribution (6.22) – (6.25) based on the variational Bayes theorem by ignoring its own contribution to the expectation. In the evaluation of the variational expectation of $p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y)$ we use the generic

superscripts $h, j, k, l, m \in \mathbb{N}$ in lieu of the step-wise updated superscripts i and $i + 1$ to derive the general result and subsequently replace h, j, k, l, m by the appropriate iterative indices.

○ **Evaluation of the variational expectation of $p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y)$**

We have by definition

$$\begin{aligned} \langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(l)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_{1:T})} \\ = \langle \ln p(y_{1:T}, x_{1:T} | a, \lambda_x, b, \lambda_y) p(a, \lambda_x, b, \lambda_y) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(l)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_{1:T})} \end{aligned} \quad (6.26)$$

Under the assumption of a factorized prior distribution

$$p(a, \lambda_x, b, \lambda_y) = p(a)p(\lambda_x)p(b)p(\lambda_y) \quad (6.27)$$

and with the likelihood

$$p(y_{1:T}, x_{1:T} | a, \lambda_x, b, \lambda_y) = p(x_1) \prod_{t=2}^T p(x_t | x_{t-1}, a, \lambda_x) \prod_{t=1}^T p(y_t | x_t, b, \lambda_y) \quad (6.28)$$

we obtain

$$\begin{aligned} \langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(l)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_{1:T})} \\ = \langle \ln p(x_1) \rangle_{q^{(n)}(x_1)} \\ + \sum_{t=2}^T \langle \ln p(x_t | x_{t-1}, a, \lambda_x) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(n)}(x_{t-1:t})} \\ + \sum_{t=1}^T \langle \ln p(y_t | x_t, b, \lambda_y) \rangle_{q^{(j)}(b)q^{(k)}(\lambda_y)q^{(n)}(x_t)} \\ + \langle \ln p(a) \rangle_{q^{(j)}(a)} \\ + \langle \ln p(\lambda_x) \rangle_{q^{(k)}(\lambda_x)} \\ + \langle \ln p(b) \rangle_{q^{(l)}(b)} \\ + \langle \ln p(\lambda_y) \rangle_{q^{(m)}(\lambda_y)} \end{aligned} \quad (6.29)$$

Next, substituting the respective forms for the probability density functions, we obtain

$$\begin{aligned} \langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(l)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_{1:T})} \\ = \langle \ln \left(\left(\frac{\lambda_{x_1}}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\lambda_{x_1}}{2} (x_1 - \mu_{x_1})^2 \right) \right) \rangle_{q^{(n)}(x_1)} \\ + \sum_{t=2}^T \langle \ln \left(\left(\frac{\lambda_x}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\lambda_x}{2} (x_t - ax_{t-1})^2 \right) \right) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(n)}(x_{t-1:t})} \\ + \sum_{t=1}^T \langle \ln \left(\left(\frac{\lambda_y}{2\pi} \right)^{\frac{1}{2}} \exp \left(-\frac{\lambda_y}{2} (y_t - bx_t)^2 \right) \right) \rangle_{q^{(h)}(x_t)q^{(j)}(b)q^{(k)}(\lambda_y)} \end{aligned}$$

$$\begin{aligned}
& + \langle \ln \left(\left(\frac{1}{2\pi\sigma_a^2} \right)^{\frac{1}{2}} \exp \left(-\frac{1}{2\sigma_a^2} (a - \mu_a)^2 \right) \right) \rangle_{q^{(j)}(a)} \\
& + \langle \ln \left(\frac{1}{\Gamma(a_{\lambda_x})} \frac{1}{(b_{\lambda_x})^{a_{\lambda_x}}} \lambda_x^{a_{\lambda_x}-1} \exp \left(-\frac{\lambda_x}{b_{\lambda_x}} \right) \right) \rangle_{q^{(k)}(\lambda_x)} \\
& + \langle \ln \left(\left(\frac{1}{2\pi\sigma_b^2} \right)^{\frac{1}{2}} \exp \left(-\frac{1}{2\sigma_b^2} (b - \mu_b^{(i)})^2 \right) \right) \rangle_{q^{(l)}(b)} \\
& + \langle \ln \left(\frac{1}{\Gamma(a_{\lambda_y})} \frac{1}{(b_{\lambda_y})^{a_{\lambda_y}}} \lambda_y^{a_{\lambda_y}-1} \exp \left(-\frac{\lambda_y}{b_{\lambda_y}} \right) \right) \rangle_{q^{(m)}(\lambda_y)}
\end{aligned} \tag{6.30}$$

Evaluation of the logarithmic expressions then yields

$$\begin{aligned}
& \langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(l)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_{1:T})} \\
& = \frac{1}{2} \ln \left(\frac{\lambda_{x_1}}{2\pi} \right) - \frac{1}{2} \langle \lambda_{x_1} (x_t - \mu_{x_1})^2 \rangle_{q^{(n)}(x_1)} \\
& + \frac{T}{2} \langle \ln \left(\frac{\lambda_x}{2\pi} \right) \rangle_{q^{(i)}(\lambda_x)} - \frac{1}{2} \sum_{t=2}^T \langle \lambda_x (x_t - a x_{t-1})^2 \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(n)}(x_{t-1:t})} \\
& + \frac{T}{2} \langle \ln \left(\frac{\lambda_y}{2\pi} \right) \rangle_{q^{(i)}(\lambda_y)} - \frac{1}{2} \sum_{t=1}^T \langle \lambda_y (y_t - b x_t)^2 \rangle_{q^{(j)}(b)q^{(k)}(\lambda_y)q^{(n)}(x_t)} \\
& - \langle \frac{1}{2\sigma_a^2} (a - \mu_a)^2 \rangle_{q^{(j)}(a)} - \frac{1}{2} \ln 2\pi\sigma_a^2 \\
& - \ln \left(\Gamma(a_{\lambda_x}) \right) - \ln \left((b_{\lambda_x})^{a_{\lambda_x}} \right) + (a_{\lambda_x} - 1) \langle \ln \lambda_x \rangle_{q^{(k)}(\lambda_x)} - \frac{1}{b_{\lambda_x}} \langle \lambda_x \rangle_{q^{(k)}(\lambda_x)} \\
& - \langle \frac{1}{2\sigma_b^2} (b - \mu_b)^2 \rangle_{q^{(l)}(b)} - \frac{1}{2} \ln 2\pi\sigma_b^2 \\
& - \ln \left(\Gamma(a_{\lambda_y}) \right) - \ln \left((b_{\lambda_y})^{a_{\lambda_y}} \right) + (a_{\lambda_y} - 1) \langle \ln \lambda_y \rangle_{q^{(m)}(\lambda_y)} - \frac{1}{b_{\lambda_y}} \langle \lambda_y \rangle_{q^{(m)}(\lambda_y)}
\end{aligned} \tag{6.31}$$

Finally evaluation of the quadratic terms yields

$$\begin{aligned}
& \langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(l)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_{1:T})} \\
& = \frac{1}{2} \ln \left(\frac{\lambda_{x_1}}{2\pi} \right) - \frac{1}{2} \lambda_{x_1} \left(\langle x_1^2 \rangle_{q^{(h)}(x_1)} - 2 \langle x_1 \rangle_{q^{(h)}(x_1)} \mu_{x_1} + \mu_{x_1}^2 \right) \\
& + \frac{T}{2} \langle \ln \left(\frac{\lambda_x}{2\pi} \right) \rangle_{q^{(k)}(\lambda_x)} - \frac{1}{2} \langle \lambda_x (\sum_{t=2}^T x_t^2 - 2a \sum_{t=2}^T x_{t-1} x_t + a^2 \sum_{t=2}^T x_{t-1}^2) \rangle_{q^{(j)}(a)q^{(k)}(\lambda_x)q^{(n)}(x_{t-1:t})} \\
& + \frac{T}{2} \langle \ln \left(\frac{\lambda_y}{2\pi} \right) \rangle_{q^{(m)}(\lambda_y)} - \frac{1}{2} \langle \lambda_y (\sum_{t=1}^T y_t^2 - 2b \sum_{t=1}^T y_t x_t + b^2 \sum_{t=1}^T x_t^2) \rangle_{q^{(j)}(b)q^{(m)}(\lambda_y)q^{(n)}(x_t)} \\
& - \frac{1}{2\sigma_a^2} \langle a^2 - 2a\mu_a + \mu_a^2 \rangle_{q^{(j)}(a)} - \frac{1}{2} \ln 2\pi\sigma_a^2 \\
& - \ln \left(\Gamma(a_{\lambda_x}) \right) - a_{\lambda_x} \ln(b_{\lambda_x}) + (a_{\lambda_x} - 1) \langle \ln \lambda_x \rangle_{q^{(k)}(\lambda_x)} - \frac{1}{b_{\lambda_x}} \langle \lambda_x \rangle_{q^{(k)}(\lambda_x)}
\end{aligned}$$

$$\begin{aligned}
& -\frac{1}{2\sigma_b^2} \langle b^2 - 2b\mu_b + \mu_b^2 \rangle_{q^{(i)}(b)} - \frac{1}{2} \ln 2\pi\sigma_b^2 \\
& - \ln \left(\Gamma(a_{\lambda_y}) \right) - a_{\lambda_y} \ln(b_{\lambda_y}) + (a_{\lambda_y} - 1) \langle \ln \lambda_y \rangle_{q^{(m)}(\lambda_y)} - \frac{1}{b_{\lambda_y}} \langle \lambda_y \rangle_{q^{(m)}(\lambda_y)}
\end{aligned} \tag{6.32}$$

□

○ Application of the Variational Bayes Inference Theorem

Based on (6.32), we now apply

$$p(\vartheta_i) \propto \exp\left(\int q(\vartheta_{\setminus i}) \ln p(y, \vartheta) d\vartheta_{\setminus i}\right) = \exp\left(\langle \ln(p(y, \vartheta)) \rangle_{q(\vartheta_{\setminus i})}\right) \tag{6.33}$$

for each variational parameter distribution $q(a)$, $q(\lambda_x)$, $q(b)$, and $q(\lambda_y)$. Therefore, starting with (6.32), we ignore the expectations under the respective variational distribution and we subsume all terms independent of the variable under consideration in the proportionality constant.

○ Derivation of an update equation for $q(a)$

Setting $k, n = i$ we have

$$\ln q^{(i+1)}(a) \propto -\frac{1}{2} \left(\sum_{t=2}^T \lambda_x x_t^2 - 2a \sum_{t=2}^T \lambda_x x_{t-1} x_t + a^2 \sum_{t=2}^T \lambda_x x_{t-1}^2 \right)_{q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} - \frac{1}{2\sigma_a^2} (a^2 - 2a\mu_a + \mu_a^2) \tag{6.34}$$

Partitioning the expectations, we obtain

$$\ln q(a) \propto -\frac{1}{2} \left(\sum_{t=2}^T \langle \lambda_x x_t^2 \rangle_{q^{(i)}(\lambda_x)q^{(i)}(x_t)} - 2a \sum_{t=2}^T \langle \lambda_x x_{t-1} x_t \rangle_{q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} + a^2 \sum_{t=2}^T \langle \lambda_x x_{t-1}^2 \rangle_{q^{(i)}(\lambda_x)q^{(i)}(x_{t-1})} \right) - \frac{1}{2\sigma_a^2} (a^2 - 2a\mu_a + \mu_a^2) \tag{6.35}$$

Rewriting the above as a quadratic form in a and ignoring terms independent of a yields

$$\ln q(a) \propto -\frac{1}{2} \left(\sum_{t=2}^T \langle \lambda_x x_{t-1}^2 \rangle_{q^{(i)}(\lambda_x)q^{(i)}(x_{t-1})} + \frac{1}{\sigma_a^2} \right) a^2 + \left(\sum_{t=2}^T \langle \lambda_x x_{t-1} x_t \rangle_{q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} + \frac{\mu_a}{\sigma_a^2} \right) a \tag{6.36}$$

Using

$$\langle \lambda_x x_{t-1}^2 \rangle_{q^{(i)}(\lambda_x)q^{(i)}(x_{t-1})} = \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} = a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} \tag{6.37}$$

and

$$\langle \lambda_x x_{t-1} x_t \rangle_{q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} = \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} = a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} \tag{6.38}$$

evaluation of the remaining expectations in terms of the parameters of $q^{(i)}(\lambda_x)$ then yields

$$\ln q(a) \propto -\frac{1}{2} \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} + \frac{1}{\sigma_a^2} \right) a^2 + \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} + \frac{\mu_a}{\sigma_a^2} \right) a \tag{6.39}$$

According to the completing-the-square theorem (Supplement Section 2),

$$\exp\left(-\frac{1}{2} \alpha a^2 + \beta a\right) \propto N(a; \alpha^{-1} \beta, \alpha^{-1}) \tag{6.40}$$

we thus have

$$q^{(i+1)}(a) \propto N\left(a; \mu_a^{(i+1)}, \sigma_a^{2(i+1)}\right) \quad (6.41)$$

where

$$\sigma_a^{2(i+1)} := \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} + \frac{1}{\sigma_a^2} \right)^{-1} \quad (6.42)$$

and

$$\mu_a^{(i+1)} := \sigma_a^{2(i+1)} \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} + \frac{\mu_a}{\sigma_a^2} \right) \quad (6.43)$$

□

○ Derivation of an update equation for $q(\lambda_x)$

From (6.32), using the expectation partitioning as above, reordering in terms of $\ln \lambda_x$ and λ_x and setting $j = i + 1$ and $n = i$ we have

$$\ln q^{(i+1)}(\lambda_x) \propto \left(\frac{T}{2} + a_{\lambda_x} - 1 \right) \ln \lambda_x - \left(\frac{1}{2} \left(\sum_{t=2}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} - 2 \sum_{t=2}^T \langle a x_{t-1} x_t \rangle_{q^{(i+1)}(a) q^{(i)}(x_{t-1:t})} + \sum_{t=2}^T \langle a^2 x_{t-1}^2 \rangle_{q^{(i+1)}(a) q^{(i)}(x_{t-1:t})} \right) + \frac{1}{b_{\lambda_x}} \right) \lambda_x \quad (6.44)$$

Taking the exponential on both sides then yields

$$q^{(i+1)}(\lambda_x) \propto \lambda_x^{\left(\frac{T}{2} + a_{\lambda_x} - 1 \right)} \exp \left(- \left(\frac{1}{b_{\lambda_x}} + \frac{1}{2} \left(\sum_{t=2}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} - 2 \sum_{t=2}^T \langle a x_{t-1} x_t \rangle_{q^{(i+1)}(a) q^{(i)}(x_{t-1:t})} + \sum_{t=2}^T \langle a^2 x_{t-1}^2 \rangle_{q^{(i+1)}(a) q^{(i)}(x_{t-1:t})} \right) \right) \lambda_x \right) \quad (6.45)$$

Using

$$\langle a x_{t-1} x_t \rangle_{q^{(i+1)}(a) q^{(i)}(x_{t-1:t})} = \langle a \rangle_{q^{(i+1)}(a)} \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} = \mu_a^{(i+1)} \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} \quad (6.46)$$

and

$$\langle a^2 x_{t-1}^2 \rangle_{q^{(i+1)}(a) q^{(i)}(x_{t-1})} = \langle a^2 \rangle_{q^{(i+1)}(a)} \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} = \left(\sigma_a^{2(i+1)} + \mu_a^{(i+1)^2} \right) \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} \quad (6.47)$$

evaluation of the remaining expectations in terms of the parameters of $q^{(i+1)}(a)$ then yields

$$q^{(i+1)}(\lambda_x) \propto \lambda_x^{\left(\frac{T}{2} + a_{\lambda_x} - 1 \right)} \exp \left(- \left(\frac{1}{b_{\lambda_x}} + \frac{1}{2} \left(\sum_{t=2}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} - 2 \sum_{t=2}^T \mu_a^{(i+1)} \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} + \sum_{t=2}^T \left(\sigma_a^{2(i+1)} + \mu_a^{(i+1)^2} \right) \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} \right) \right) \lambda_x \right) \quad (6.48)$$

Up to a normalization constant, $q^{(i+1)}(\lambda_x)$ is thus given by a Gamma distribution

$$q^{(i+1)}(\lambda_x) \propto \lambda^{a_{\lambda_x}^{(i+1)} - 1} \exp \left(- \frac{\lambda_x}{b_{\lambda_x}^{(i+1)}} \right) \quad (6.49)$$

with

$$a_{\lambda_x}^{(i+1)} := \frac{T}{2} + a_{\lambda_x} \quad (6.50)$$

and

$$b_{\lambda_x}^{(i+1)} := \left(\frac{1}{b_{\lambda_x}} + \frac{1}{2} \left(\sum_{t=2}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} - 2 \sum_{t=2}^T \mu_a^{(i+1)} \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} + \left(\sigma_a^{2(i+1)} + \mu_a^{(i+1)^2} \right) \sum_{t=2}^T \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} \right) \right)^{-1} \quad (6.51)$$

□

○ Derivation of an update equation for $q(b)$

Setting $m = n = i$ we have

$$\ln q(b) \propto -\frac{1}{2} \langle \lambda_y (\sum_{t=1}^T y_t^2 - 2b \sum_{t=1}^T y_t x_t + b^2 \sum_{t=1}^T x_t^2) \rangle_{q^{(i)}(\lambda_y) q^{(i)}(x_{t-1:t})} - \frac{1}{2\sigma_b^2} (b^2 - 2b\mu_b + \mu_b^2) \quad (6.52)$$

Partitioning the expectations, we obtain

$$\ln q(b) \propto -\frac{1}{2} \langle \lambda_y \rangle_{q^{(i)}(\lambda_y)} \sum_{t=1}^T y_t y_t - 2b \sum_{t=1}^T y_t \langle \lambda_y x_t \rangle_{q^{(i)}(\lambda_y) q^{(i)}(x_t)} + b^2 \sum_{t=1}^T \langle \lambda_y x_t^2 \rangle_{q^{(i)}(\lambda_y) q^{(i)}(x_t)} - \frac{1}{2\sigma_b^2} (b^2 - 2b\mu_b + \mu_b^2) \quad (6.53)$$

Rewriting the above as a quadratic form in b and ignoring terms independent of b yields

$$\ln q(b) \propto -\frac{1}{2} \left(\sum_{t=1}^T \langle \lambda_y x_t^2 \rangle_{q^{(i)}(\lambda_y) q^{(i)}(x_t)} + \frac{1}{\sigma_b^2} \right) b^2 + \left(\sum_{t=1}^T y_t \langle \lambda_y x_t \rangle_{q^{(i)}(\lambda_y) q^{(i)}(x_t)} + \frac{\mu_b}{\sigma_b^2} \right) b \quad (6.54)$$

Using

$$\langle \lambda_y x_t^2 \rangle_{q^{(i)}(\lambda_y) q^{(i)}(x_t)} = \langle \lambda_y \rangle_{q^{(i)}(\lambda_y)} \langle x_t^2 \rangle_{q^{(i)}(x_t)} = a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \langle x_t^2 \rangle_{q^{(i)}(x_t)} \quad (6.55)$$

and

$$\langle \lambda_y x_t \rangle_{q^{(i)}(\lambda_y) q^{(i)}(x_t)} = \langle \lambda_y \rangle_{q^{(i)}(\lambda_y)} \langle x_t \rangle_{q^{(i)}(x_t)} = a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \langle x_t \rangle_{q^{(i)}(x_t)} \quad (6.56)$$

evaluation of the remaining expectations in terms of the parameters of $q^{(i)}(\lambda_y)$ then yields

$$\ln q(b) \propto -\frac{1}{2} \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} + \frac{1}{\sigma_b^2} \right) b^2 + \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T y_t \langle x_t \rangle_{q^{(i)}(x_t)} + \frac{\mu_b}{\sigma_b^2} \right) b \quad (6.57)$$

According to the completing-the-square theorem,

$$\exp \left(-\frac{1}{2} \alpha b^2 + \beta b \right) \propto N(b; \alpha^{-1} \beta, \alpha^{-1}) \quad (6.58)$$

we thus have

$$q^{(i+1)}(b) \propto N(b; \mu_a^{(i+1)}, \sigma_b^{2(i+1)}) \quad (6.59)$$

where

$$\sigma_b^{2(i+1)} := \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} + \frac{1}{\sigma_b^2} \right)^{-1} \quad (6.60)$$

and

$$\mu_b^{(i+1)} := \sigma_b^{2(i+1)} \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T y_t \langle x_t \rangle_{q^{(i)}(x_t)} + \frac{\mu_b}{\sigma_b^2} \right) \quad (6.61)$$

□

○ Derivation of an update equation for $q(\lambda_y)$

From (6.32), using the expectation partitioning as above, re-ordering in terms of $\ln \lambda_y$ and λ_y and setting $l = i + 1$ and $n = i$ we have

$$\ln q^{(i+1)}(\lambda_y) \propto \left(\frac{T}{2} + a_{\lambda_y} - 1 \right) \ln \lambda_y - \left(\frac{1}{2} \left(\sum_{t=1}^T y_t^2 - 2 \sum_{t=1}^T y_t \langle b x_t \rangle_{q^{(i+1)}(b) q^{(i)}(x_t)} + \sum_{t=1}^T \langle b^2 x_t^2 \rangle_{q^{(i+1)}(b) q^{(i)}(x_t)} \right) + \frac{1}{b_{\lambda_y}} \right) \lambda_y \quad (6.62)$$

Taking the exponential on both sides then yields

$$q^{(i+1)}(\lambda_y) \propto \lambda_y^{\left(\frac{T}{2} + a_{\lambda_y} - 1\right)} \exp\left(-\left(\frac{1}{2}\left(\sum_{t=1}^T y_t^2 - 2 \sum_{t=1}^T y_t \langle b x_t \rangle_{q^{(i+1)}(b) q^{(i)}(x_t)} + \sum_{t=1}^T \langle b^2 x_t^2 \rangle_{q^{(i+1)}(b) q^{(i)}(x_t)}\right) + \frac{1}{b_{\lambda_y}}\right) \lambda_y\right) \quad (6.63)$$

Using

$$\langle b x_t \rangle_{q^{(i+1)}(b) q^{(i)}(x_t)} = \langle b \rangle_{q^{(i+1)}(b)} \langle x_t \rangle_{q^{(i)}(x_t)} = \mu_b^{(i+1)} \langle x_t \rangle_{q^{(i)}(x_t)} \quad (6.64)$$

and

$$\langle b^2 x_t^2 \rangle_{q^{(i+1)}(b) q^{(i)}(x_t)} = \langle b^2 \rangle_{q^{(i+1)}(b)} \langle x_t^2 \rangle_{q^{(i)}(x_t)} = \left(\sigma_b^{2(i+1)} + \mu_b^{(i+1)2}\right) \langle x_t^2 \rangle_{q^{(i)}(x_t)} \quad (6.65)$$

evaluation of the remaining expectations in terms of the parameters of $q^{(i)}(b)$ then yields

$$q^{(i+1)}(\lambda_y) \propto \lambda_y^{\left(\frac{T}{2} + a_{\lambda_y} - 1\right)} \exp\left(-\left(\frac{1}{2}\left(\sum_{t=1}^T y_t^2 - 2 \mu_b^{(i+1)} \sum_{t=1}^T y_t \langle x_t \rangle_{q^{(i)}(x_t)} + \left(\sigma_b^{2(i+1)} + \mu_b^{(i+1)2}\right) \sum_{t=1}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)}\right) + \frac{1}{b_{\lambda_y}}\right) \lambda_y\right) \quad (6.66)$$

Up to a normalization constant, $q^{(i+1)}(\lambda_y)$ is thus given by a Gamma distribution

$$q^{(i+1)}(\lambda_y) \propto \lambda_y^{a_{\lambda_y}^{(i+1)} - 1} \exp\left(-\frac{\lambda_y}{b_{\lambda_y}^{(i+1)}}\right) \quad (6.67)$$

with

$$a_{\lambda_y}^{(i+1)} := \frac{T}{2} + a_{\lambda_y} \quad (6.68)$$

and

$$b_{\lambda_y}^{(i+1)} := \left(\frac{1}{b_{\lambda_y}} + \frac{1}{2}\left(\sum_{t=1}^T y_t^2 - 2 \mu_b^{(i+1)} \sum_{t=1}^T y_t \langle x_t \rangle_{q^{(i)}(x_t)} + \left(\sigma_b^{2(i+1)} + \mu_b^{(i+1)2}\right) \sum_{t=1}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)}\right)\right)^{-1} \quad (6.69)$$

□

In summary, we have obtained a set of iterative update equations for the variational distribution parameters which can be initialized using the parameters of the marginal distributions (6.4) – (6.7) and is shown in pseudocode form in Table 4. Notably, most update equations involve sums over expectations of the unobserved variables $x_{1:T}$ under the variational distribution $q^{(i)}(x_{1:T})$. The derivation of these expectations in parameterized form is the topic of Section 6.3.1.5 of this supplementary material.

% initialization of the variational distribution parameters to the parameters of the prior distributions

$$\sigma_a^{2(0)} := \sigma_a^2$$

$$\mu_a^{(0)} := \mu_a$$

$$a_{\lambda_x}^{(0)} := a_{\lambda_x}$$

$$b_{\lambda_x}^{(0)} := b_{\lambda_x}$$

$$\sigma_b^{2(0)} := \sigma_b^2$$

$$\mu_b^{(0)} := \mu_b$$

$$a_{\lambda_y}^{(0)} := a_{\lambda_y}$$

$$b_{\lambda_y}^{(i)} := b_{\lambda_y}$$

% iterative update equations for the variational distribution parameters

for $i = 0, 1, 2 \dots$

$$\sigma_a^{2(i+1)} := \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} + \frac{1}{\sigma_a^2} \right)^{-1}$$

$$\mu_a^{(i+1)} := \sigma_a^{2(i+1)} \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} + \frac{\mu_a}{\sigma_a^2} \right)$$

$$a_{\lambda_x}^{(i+1)} := \frac{T}{2} + a_{\lambda_x}$$

$$b_{\lambda_x}^{(i+1)} := \left(\frac{1}{b_{\lambda_x}} + \frac{1}{2} \left(\sum_{t=2}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} - 2 \sum_{t=2}^T \mu_a^{(i+1)} \langle x_{t-1} x_t \rangle_{q^{(i)}(x_{t-1:t})} + \left(\sigma_a^{2(i+1)} + \mu_a^{(i+1)^2} \right) \sum_{t=2}^T \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} \right) \right)^{-1}$$

$$\sigma_b^{2(i+1)} := \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} + \frac{1}{\sigma_b^2} \right)^{-1}$$

$$\mu_b^{(i+1)} := \sigma_b^{2(i+1)} \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T y_t \langle x_t \rangle_{q^{(i)}(x_t)} + \frac{\mu_b}{\sigma_b^2} \right)$$

$$a_{\lambda_y}^{(i+1)} := \frac{T}{2} + a_{\lambda_y}$$

$$b_{\lambda_y}^{(i+1)} := \left(\frac{1}{b_{\lambda_y}} + \frac{1}{2} \left(\sum_{t=1}^T y_t^2 - 2 \mu_b^{(i+1)} \sum_{t=1}^T y_t \langle x_t \rangle_{q^{(i)}(x_t)} + \left(\sigma_b^{2(i+1)} + \mu_b^{(i+1)^2} \right) \sum_{t=1}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} \right) \right)^{-1}$$

Table 4. Variational parameter update equations for the unobserved **variables** a, b, λ_x and λ_y .

6.3 Evaluation of $q^{(i+1)}(x_{1:T})$

We first validate the relationship between the proportionality statements of equations (28) and (29) of the tutorial:

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(y_{1:T}, x_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right) \Leftrightarrow q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(x_{1:T} | y_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right) \quad (6.70)$$

○ Derivation of the equivalence (6.70)

We have

$$\begin{aligned} q^{(i+1)}(x_{1:T}) &\propto \exp\left(\langle \ln p(y_{1:T}, x_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln(p(y_{1:T}, x_{1:T} | \theta) p(\theta)) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln p(y_{1:T}, x_{1:T} | \theta) + \ln p(\theta) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)} + \langle \ln p(\theta) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)}\right) \exp\left(\langle \ln p(\theta) \rangle_{q^{(i+1)}(\theta)}\right) \end{aligned} \quad (6.71)$$

The latter term in the product on the right-hand side of (6.71) is independent of $x_{1:T}$ and may thus be regarded as a (further) positive multiplicative constant of the probability density governing the $x_{1:T}$. We thus have

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)}\right) \quad (6.72)$$

We further have

$$\begin{aligned} q^{(i+1)}(x_{1:T}) &\propto \exp\left(\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln(p(x_{1:T} | y_{1:T}, \theta) p(y_{1:T} | \theta)) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln p(x_{1:T} | y_{1:T}, \theta) + \ln p(y_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln p(x_{1:T} | y_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)} + \langle \ln p(y_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)}\right) \\ &= \exp\left(\langle \ln p(x_{1:T} | y_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right) \exp(\langle \ln p(y_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)}) \end{aligned} \quad (6.72a)$$

Again, the latter term in the product on the right-hand side of (6.73) is independent of $x_{1:T}$ and may thus be regarded as yet another positive multiplicative constant of the probability density governing the $x_{1:T}$. We thus have

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(x_{1:T} | y_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right) \quad (6.73)$$

□

As discussed in Section 4.2.2, inference of $\exp\left(\langle \ln p(x_{1:T} | y_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right)$ can be achieved by capitalizing on classic state space model inference algorithms. As in the tutorial, we thus first review inference of $p_\theta(x_{1:T} | y_{1:T})$,

i.e. the evaluation of $p(x_{1:T}|y_{1:T})$ for known, non-random values of $\theta := (a, \sigma_x^2, b, \sigma_y^2)$. To this end, we first verify equation (31) of the tutorial, i.e. the re-parameterization of $p(x_{1:T}|y_{1:T})$ in terms of a product the ratio between its pairwise conditional marginal distributions $p(x_{t-1}, x_t|y_{1:T})$ and its conditional marginal distributions $p(x_{t-1}|y_{1:T})$:

$$p(x_{1:T}|y_{1:T}) = p(x_1|y_{1:T}) \prod_{t=2}^T \frac{p(x_{t-1}, x_t|y_{1:T})}{p(x_{t-1}|y_{1:T})} \quad (6.74)$$

◦ Derivation of the equivalence (6.74)

From the definition of conditional probability, we have

$$\begin{aligned} p(x_{1:T}|y_{1:T}) &= p(x_T|x_{1:T-1}, y_{1:T})p(x_{T-1}|x_{1:T-2}, y_{1:T}) \cdots p(x_2|x_1, y_{1:T}) p(x_1|y_{1:T}) \\ &= p(x_1|y_{1:T}) \prod_{t=2}^T p(x_t|x_{1:t-1}, y_{1:T}) \end{aligned} \quad (6.75)$$

From the defining properties of the LGSSM, we have that given x_{t-1} , the random variable x_t is conditionally independent of all other latent variables $x_{1:t-2}$ and $x_{t+1:T}$ ($t = 3, \dots, T-1$). We may thus discard the variables $x_{1:t-2}$ from each term in the product of the right-hand side of equation (6.75). Likewise, the random variable x_1 was assumed to be conditionally independent of $x_{2:T}$ (or in other words, defined by its marginal distribution). We thus have

$$p(x_{1:T}|y_{1:T}) = p(x_1|y_{1:T}) \prod_{t=2}^T p(x_t|x_{t-1}, y_{1:T}) = p(x_1|y_{1:T}) \prod_{t=2}^T \frac{p(x_{t-1}, x_t|y_{1:T})}{p(x_{t-1}|y_{1:T})} \quad (6.76)$$

□

Next, we provide a general review of the Kalman-Rauch-Tung Striebel Algorithm. Because, in contrast to the scenario discussed in the tutorial, standard treatments of LGSSMs consider multivariate latent and observed variables, the following treatment uses the multivariate form of the LGSSM.

6.3.1 The Kalman-Rauch-Tung-Striebel Smoothing Algorithm

For LGSSMs, the evaluation of conditional distributions over the variables $x_{1:T}$ given an observed data sequence $y_{1:T}$ and a known parameter θ corresponds to the “state space model inference” problem for which a variety of solutions exist in the literature (for a comprehensive review see e.g. (Briers, Doucet, & Maskell, 2010)). The most popular solution for inferring the distributions of $x_{1:T}$ is perhaps the KRTS smoother, which represents the combination of the classic Kalman Filter for state space models (Kalman, 1960) with a backward algorithm (Johari, Desabrais, Rauch, Striebel, & Tung, 1965). In brief, the KRTS smoother is a recursive algorithm that yields the expectation and covariance parameters of the conditional unobserved variable distributions in matrix product form.

In contrast to the tutorial, we here consider the general form of a “multivariate” LGSSM. To this end, we denote the dimensionality of the unobserved variables by $k \in \mathbb{N}$ and the dimensionality of the observed variable by $p \in \mathbb{N}$ and rewrite expression (5) of the tutorial as

$$x_t = Ax_{t-1} + \varepsilon_t \quad (x_t \in \mathbb{R}^k, A \in \mathbb{R}^{k \times k}, \varepsilon_t \sim N(\varepsilon_t; 0, \Sigma_x), 0 \in \mathbb{R}^k, \Sigma_x \in \mathbb{R}^{k \times k}, \Sigma_x > 0) \quad (6.77)$$

$$y_t = Bx_t + \eta_t \quad (y_t \in \mathbb{R}^p, B \in \mathbb{R}^{p \times k}, \eta_t \sim N(\eta_t; 0, \Sigma_y), 0 \in \mathbb{R}^p, \Sigma_y \in \mathbb{R}^{p \times p}, \Sigma_y > 0)$$

for $t = 2, \dots, T$ and $t = 1, \dots, T$, respectively, and with the assumption of independent dynamics and observation noise, that is, $\mathbb{C}(\varepsilon_t, \eta_t) = 0 \in \mathbb{R}^{k \times p}$. It is central for the understanding of LGSSMs that these two equations specify a multivariate joint Gaussian distribution over the unobserved variables $x_1, \dots, x_T$ and the observed variables $y_1, \dots, y_T$. Specifically, due to the Gaussian properties of ε_t and η_t , the equations of (6.77) specify the following conditional probability distributions over unobserved and observed variables x_t and y_t , respectively

$$p(x_t|x_{t-1}) = N(x_t; Ax_{t-1}, \Sigma_x), t = 2, \dots, T \text{ and } p(y_t|x_t) = N(y_t; Bx_t, \Sigma_y), t = 1, \dots, T \quad (6.78)$$

The specification of the joint distribution for the LGSSM is completed by a further Gaussian distribution over x_1 with expectation parameter $\mu_1 \in \mathbb{R}^k$ and covariance $\Sigma_1 \in \mathbb{R}^{k \times k}$

$$p(x_1) = N(x_1; \mu_1, \Sigma_1) \quad (6.79)$$

Using the “colon notation” abbreviations

$$x_{1:T} := (x_1, \dots, x_T) \text{ and } y_{1:T} := (y_1, \dots, y_T) \quad (6.80)$$

and parameter set definition $\theta := \{\mu_1, \Sigma_1, A, \Sigma_x, B, \Sigma_y\}$ then allows for writing the joint distribution over unobserved and observed variables as

$$p_\theta(x_{1:T}, y_{1:T}) = p_{\mu_1, \Sigma_1}(x_1) \prod_{t=2}^T p_{A, \Sigma_x}(x_t|x_{t-1}) \prod_{t=1}^T p_{B, \Sigma_y}(y_t|x_t) \quad (6.81)$$

As discussed in the tutorial, expression (6.81) states that the joint distribution over all variables $x_{1:T}$ and $y_{1:T}$ of the LGSSM is given by the product of conditional Gaussian distributions over $x_{1:T}$ and $y_{1:T}$, respectively. Because the product of Gaussian probability densities is again a Gaussian probability density, (6.81) hence specifies a Gaussian joint distribution by means of its factorization properties.

The Kalman-Rauch-Tung-Striebel (KRTS) smoothing algorithm, as sketched herein, is an algorithm that, given a parameterized $2T$ -dimensional Gaussian joint probability $p(x_{1:T}, y_{1:T})$ over random variables $x_{1:T}$ and $y_{1:T}$ with specific factorization properties and parameterized by the parameter vector θ , returns the parameters of the T -dimensional conditional distribution $p(x_{1:T}|y_{1:T})$. However, as outlined in the tutorial, the parameters of $p(x_{1:T}|y_{1:T})$ are returned not in the form of a T -dimensional expectation parameter $\mu_{x_{1:T}|y_{1:T}}$ and $T \times T$ -dimensional covariance parameter $\Sigma_{x_{1:T}|y_{1:T}}$, but as the parameters of the conditional “marginal” distributions $p_\theta(x_t|y_{1:T})$ for $t = 1, \dots, T$ and the conditional “joint” distributions $p(x_{t-1}, x_t|y_{1:T})$. Because, as shown above, by the conditional independence properties of the LGSSM, we have

$$p(x_{1:T}|y_{1:T}) = p(x_1|y_{1:T}) \prod_{t=2}^T \frac{p(x_{t-1}, x_t|y_{1:T})}{p(x_{t-1}|y_{1:T})} \quad (6.82)$$

these two parameterizations are equivalent. Following the classic treatment of state space model inference, in this Section, we first review the Kalman filter, a recursive algorithm to obtain the parameters of the distributions $p(x_t|y_{1:t})$ in Sections 6.3.1.1 and 6.3.1.2. We then turn to the KRTS smoothing algorithm in Section 6.3.1.3. Based on the results of the Kalman filter algorithm, the KRTS smoothing algorithm allows for the derivation of the parameters of the distributions $p(x_t|y_{1:T})$ in the denominator of the right-hand side of (6.83). Finally, in Section 6.3.1.4, we review the recursive evaluation of the parameters $p(x_{t-1}, x_t|y_{1:T})$ required in the numerator of the right-hand side of (6.83) and close this section with a summary in 6.3.1.5.

6.3.1.1 Inference of $p(x_t|y_{1:t})$

Inferring a probability distribution over the unobserved variable x_t given the values of all observed variables up to y_t , i.e., $p(x_t|y_{1:t})$ is performed readily due to the analytical properties of Gaussians distributions. As the joint distribution over $x_{1:T}$ and $y_{1:T}$ is Gaussian, all marginal and conditional marginal distributions are also Gaussian distributions. Further, as Gaussian distributions are characterized by their first two central moments, i.e., their expectation and covariance parameters, it is only these parameters that have to be inferred to fully characterize the filter distribution of interest

$$p(x_t|y_{1:t}) = N(x_t; \mu_{x_t|y_{1:t}}, \Sigma_{x_t|y_{1:t}}) \quad (t = 1, \dots, T) \quad (6.84)$$

The aim of the following section is to find a recursive expression for the conditional expectation and covariance parameters $\mu_{x_t|y_{1:t}} \in \mathbb{R}^k$ and $\Sigma_{x_t|y_{1:t}} \in \mathbb{R}^{k \times k}$ of (6.84) in terms of the parameters of the “preceding distribution”

$$p(x_{t-1}|y_{1:t-1}) = N(x_{t-1}; \mu_{x_{t-1}|y_{1:t-1}}, \Sigma_{x_{t-1}|y_{1:t-1}}) \quad (t = 2, \dots, T) \quad (6.85)$$

Formally, this can be achieved in two steps (in the derivation of the Kalman-filter equations, we follow (David Barber, 2012)): first, by finding the parameters of

$$p(x_t, y_t|y_{1:t-1}) = N\left(\begin{pmatrix} x_t \\ y_t \end{pmatrix}; \mu_{x_t, y_t|y_{1:t-1}}, \Sigma_{x_t, y_t|y_{1:t-1}}\right) \quad (6.86)$$

in terms of the parameters of the distribution $p(x_{t-1}, y_{t-1}|y_{1:t-1})$, and second, by conditioning $p(x_t, y_t|y_{1:t-1})$ on y_t , yielding the parameters of the distribution of interest. The first step can be achieved by considering the partitioning of the expectation and covariance parameters of $p(x_t, y_t|y_{1:t-1})$, which are given by

$$\mu_{x_t, y_t|y_{1:t-1}} = \begin{pmatrix} \mu_{x_t|y_{1:t-1}} \\ \mu_{y_t|y_{1:t-1}} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(x_t|y_{1:t-1}) \\ \mathbb{E}(y_t|y_{1:t-1}) \end{pmatrix} \in \mathbb{R}^{k+p} \quad (6.87)$$

and

$$\Sigma_{x_t, y_t|y_{1:t-1}} = \begin{pmatrix} \Sigma_{x_t, x_t|y_{1:t-1}} & \Sigma_{x_t, y_t|y_{1:t-1}} \\ \Sigma_{y_t, x_t|y_{1:t-1}} & \Sigma_{y_t, y_t|y_{1:t-1}} \end{pmatrix} = \begin{pmatrix} \mathbb{C}(x_t, x_t|y_{1:t-1}) & \mathbb{C}(x_t, y_t|y_{1:t-1}) \\ \mathbb{C}(y_t, x_t|y_{1:t-1}) & \mathbb{C}(y_t, y_t|y_{1:t-1}) \end{pmatrix} \in \mathbb{R}^{(k+p) \times (k+p)} \quad (6.88)$$

Recursive expressions of (6.87) and (6.88) in terms of the parameters of $p(x_{t-1}, y_{t-1}|y_{1:t-2})$ can be obtained by means of the linear transformation theorem for Gaussian distributions, yielding (see below for details)

$$\mu_{x_t, y_t|y_{1:t-1}} = \begin{pmatrix} A\mu_{x_{t-1}|y_{1:t-1}} \\ BA\mu_{x_{t-1}|y_{1:t-1}} \end{pmatrix} \in \mathbb{R}^{k+p} \quad (6.89)$$

and

$$\Sigma_{x_t, y_t|y_{1:t-1}} = \begin{pmatrix} A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x & B^T(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x) \\ B(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x)^T & B(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x)B^T + \Sigma_y \end{pmatrix} \in \mathbb{R}^{(k+p) \times (k+p)} \quad (6.90)$$

◦ Derivation of equations (6.89) and (6.90)

The first component of $\mu_{x_t, y_t|y_{1:t-1}}$ represents the expectation of the random vector x_t conditioned on the observations $y_{1:t-1}$. Based on the dynamics equation of the LGSSM it is thus obtained by a linear transformation of the expectation of the random vector x_{t-1} conditioned on the observations $y_{1:t-1}$ under the transition matrix A and additive zero-mean noise ε_t , in other words

$$\mu_{x_t|y_{1:t-1}} = \mathbb{E}(x_t|y_{1:t-1}) = \mathbb{E}(Ax_{t-1} + \varepsilon_t|y_{1:t-1}) = \mathbb{E}(Ax_{t-1}|y_{1:t-1}) = A\mathbb{E}(x_{t-1}|y_{1:t-1}) = A\mu_{x_{t-1}|y_{1:t-1}} \quad (6.91)$$

Likewise, the expectation of the observation y_t conditioned on the observations $y_{1:t-1}$ is given by the expectation of the linear transformation of x_t under the emission matrix B and additive zero-mean noise η_t ; and further, x_t is itself given by the linear transformation of x_{t-1} under A , and additive zero-mean noise ε_t in other words

$$\mu_{y_t|y_{1:t-1}} = \mathbb{E}(y_t|y_{1:t-1}) = \mathbb{E}(Bx_t + \eta_t|y_{1:t-1}) = \mathbb{E}(B(Ax_{t-1} + \varepsilon_t)|y_{1:t-1}) = B\mathbb{E}(Ax_{t-1} + \varepsilon_t|y_{1:t-1}) = BA\mathbb{E}(x_{t-1}|y_{1:t-1}) = BA\mu_{x_{t-1}|y_{1:t-1}} \quad (6.92)$$

Concatenation then yields the desired recursive expression for the expectation parameter of $p(x_t, y_t|y_{1:t-1})$

$$\mu_{x_t, y_t | y_{1:t-1}} = \begin{pmatrix} A\mu_{x_{t-1} | y_{1:t-1}} \\ BA\mu_{x_{t-1} | y_{1:t-1}} \end{pmatrix} \quad (6.93)$$

The first component of $\Sigma_{x_t, y_t | y_{1:t-1}}$ represents the covariance of the random vector x_t with itself conditioned on $y_{1:t-1}$, $\mathbb{C}(x_t, x_t | y_{1:t-1})$. As x_t is given as the linear transformation of x_{t-1} under additive zero mean noise ε_t with covariance matrix Σ_x , the Gaussian linear transformation theorem (conditioned on $y_{1:t-1}$) applies. The expression for the covariance of x_t with x_t in terms of the covariance of x_{t-1} with x_{t-1} is thus given by

$$\mathbb{C}(x_t, x_t | y_{1:t-1}) = A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x \quad (6.94)$$

Evaluating the covariance of y_t with y_t in terms of the covariance of y_{t-1} with y_{t-1} yields

$$\begin{aligned} \mathbb{C}(y_t, y_t | y_{1:t-1}) &= \mathbb{C}(Bx_t + \eta_t, Bx_t + \eta_t | y_{1:t-1}) \\ &= \mathbb{C}(Bx_t, Bx_t | y_{1:t-1}) + \Sigma_y \\ &= B\mathbb{C}(x_t, x_t | y_{1:t-1})B^T + \Sigma_y \\ &= B(A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T + \Sigma_y \end{aligned} \quad (6.95)$$

Finally, evaluation of the covariance of x_t with y_t in terms of the covariance of x_{t-1} with y_{t-1} yields

$$\mathbb{C}(x_t, y_t | y_{1:t-1}) = \mathbb{C}(Ax_{t-1} + \varepsilon_t, B(Ax_{t-1} + \varepsilon_t) + \eta_t | y_{1:t-1}) = (A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T \quad (6.96)$$

and hence also

$$\mathbb{C}(y_t, x_t | y_{1:t-1}) = (\mathbb{C}(x_t, y_t | y_{1:t-1}))^T = B(A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)^T \quad (6.97)$$

Concatenation then yields the desired recursive expression for the covariance parameter of $p(x_t, y_t | y_{1:t-1})$:

$$\Sigma_{x_t, y_t | y_{1:t-1}} = \begin{pmatrix} A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x & (A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T \\ B(A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)^T & B(A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T + \Sigma_y \end{pmatrix} \in \mathbb{R}^{(p+k) \times (p+k)} \quad (6.98)$$

□

6.3.1.2 The Kalman filter equations

The second step outlined above can be achieved by capitalizing on the Gaussian conditioning theorem (2.17), resulting in

$$\mu_{x_t | y_{1:t}} = A\mu_{x_{t-1} | y_{1:t-1}} + (A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T (B(A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T + \Sigma_y)^{-1} (y_t - BA\mu_{x_{t-1} | y_{1:t-1}}) \quad (6.99)$$

and

$$\Sigma_{x_t | y_{1:t}} = (A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x) - (A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T (B(A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x)B^T + \Sigma_y)^{-1} B(A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x) \quad (6.100)$$

The expressions (6.99) and (6.100) for $\mu_{x_t | y_{1:t}}$ and $\Sigma_{x_t | y_{1:t}}$ above can be formulated more succinctly by setting

$$P_t := \mathbb{C}(x_t, x_t | y_{1:t-1}) = A\Sigma_{x_{t-1} | y_{1:t-1}}A^T + \Sigma_x \in \mathbb{R}^{k \times k} \quad (6.101)$$

yielding

$$\mu_{x_t|y_{1:t}} = A\mu_{x_{t-1}|y_{1:t-1}} + P_t B^T (B P_t B^T + \Sigma_y)^{-1} (y_t - B A \mu_{x_{t-1}|y_{1:t-1}}) \quad (6.102)$$

and

$$\Sigma_{x_t|y_{1:t}} = P_t - P_t B^T (B P_t B^T + \Sigma_y)^{-1} B P_t \quad (6.103)$$

Finally, by defining the so-called Kalman Gain Matrix

$$K_t := \Sigma_{x_t, y_t | y_{1:t-1}} \Sigma_{y_t, y_t | y_{1:t-1}}^{-1} = P_t B^T (B P_t B^T + \Sigma_y)^{-1} \in \mathbb{R}^{k \times p} \quad (6.104)$$

the parameters may equivalently be expressed as

$$\mu_{x_t|y_{1:t}} = A\mu_{x_{t-1}|y_{1:t-1}} + K_t (y_t - B A \mu_{x_{t-1}|y_{1:t-1}}) \quad (6.105)$$

and

$$\Sigma_{x_t|y_{1:t}} = P_t - K_t B P_t = (I - K_t B) P_t \quad (6.106)$$

To avoid that the inferred covariance $\Sigma_{x_t|y_{1:t}}$, which is given by the difference of two positive-definite matrices in (6.35) and, hence, is itself positive-definite, becomes non-positive-definite due to numerical rounding errors, Bucy and Joseph (Bucy & Joseph, 1987) proposed the following modified update rule, which includes the formation of a positive-definite matrix on every inference step, which is derived below:

$$\Sigma_{x_t|y_{1:t}} = (I - K_t B) P_t = (I - K_t B) P_t (I - K_t B)^T + K_t \Sigma_y K_t^T \quad (6.107)$$

○ Derivation of equation (6.107)

Omitting subscripts to simplify the notation, we have

$$\begin{aligned} (I - KB)P(I - KB)^T + K\Sigma_y K^T &= (I - KB)P(I - B^T K^T) + K\Sigma_y K^T \\ &= (I - KB)P - (I - KB)PB^T K^T + K\Sigma_y K^T \\ &= (I - KB)P - PB^T K^T + KBPB^T K^T + K\Sigma_y K^T \\ &= (I - KB)P - PB^T K^T + K(BPB^T + \Sigma_y)K^T \\ &= (I - KB)P - PB^T K^T + PB^T (BPB^T + \Sigma_y)^{-1} (BPB^T + \Sigma_y)K^T \\ &= (I - KB)P - PB^T K^T + PB^T K^T \\ &= (I - KB)P \end{aligned} \quad (6.108)$$

□

To initialize the recursion of equations (6.105) and (6.106), we evaluate the conditional distribution $p(x_1|y_1)$ based on the parameters of the marginalization distribution $p(x_1)$ and the conditional distribution $p(y_1|x_1)$. We have

$$\mu_{x_1|y_1} := \mu_{x_1} + \Sigma_{x_1} B^T (B \Sigma_{x_1} B^T + \Sigma_y)^{-1} (y_1 - B \mu_{x_1}) \quad (6.109)$$

and

$$\Sigma_{x_1|y_1} := \Sigma_{x_1} - \Sigma_{x_1} B^T (B \Sigma_{x_1} B^T + \Sigma_y)^{-1} B \Sigma_{x_1} \quad (6.110)$$

Note the close relationship between expressions (6.109) and (6.110) and equations (6.105) and (6.106): merely replacing $A\mu_{x_{t-1}|y_{1:t-1}}$ with μ_{x_1} and $(A\Sigma_{x_{t-1}|y_{1:t-1}}A^T + \Sigma_x)$ with Σ_{x_1} in (6.109) and (6.110) results in equations

(6.105) and (6.106). Based on the derivation of equations (6.109) and (6.110) below, we may thus interpret the Kalman recursion equations as a generalization of Bayesian inference in “static” linear Gaussian models.

○ Derivation of equations (6.109) and (6.110)

By definition of the LGSSM, we have

$$p(x_1) = N(x_1; \mu_{x_1}, \Sigma_{x_1}) \text{ and } p(y_1|x_1) = N(y_1; Bx_1, \Sigma_y) \quad (6.111)$$

The expressions above specify a standard (static) Bayesian linear Gaussian model $p(x_1, y_1)$ of the form

$$p(x_1, y_1) := N\left(\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}; \begin{pmatrix} \mu_{x_1} \\ B\mu_{x_1} \end{pmatrix}, \begin{pmatrix} \Sigma_{x_1} & \Sigma_{x_1}B^T \\ B\Sigma_{x_1} & B\Sigma_{x_1}B^T + \Sigma_y \end{pmatrix}\right) \quad (6.112)$$

where

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} \mu_{x_1} \\ B\mu_{x_1} \end{pmatrix} \in \mathbb{R}^{p+k} \text{ and } \begin{pmatrix} \Sigma_{x_1} & \Sigma_{x_1}B^T \\ B\Sigma_{x_1} & B\Sigma_{x_1}B^T + \Sigma_y \end{pmatrix} \in \mathbb{R}^{(p+k) \times (p+k)} \quad (6.113)$$

(see e.g. Bishop (2006), pp. 78 - 102). From the “marginalization and conditioning theorem” of Supplement Section 2, we thus have

$$\mu_{x_1|y_1} = \mu_{x_1} + \Sigma_{x_1}B^T(B\Sigma_{x_1}B^T + \Sigma_y)^{-1}(y_1 - B\mu_{x_1}) \quad (6.114)$$

and

$$\Sigma_{x_1|y_1} = \Sigma_{x_1} - \Sigma_{x_1}B^T(B\Sigma_{x_1}B^T + \Sigma_y)^{-1}B\Sigma_{x_1} \quad (6.115)$$

□

Based on the initial equations (6.109) and (6.110), the recursive expressions (6.105) and (6.106) and the values of θ and $y_{1:T}$, we may thus infer the conditional distributions $p(x_t|y_{1:T})$ over the unobserved variables $x_{1:T}$ for $t = 2, \dots, T$. However, as discussed in the tutorial in an offline inference scenario we may also infer the conditional distributions $p(x_t|y_{1:T})$. The parametric forms of these distributions can be obtained by combining the Kalman filter recursion with a backward recursion starting from $p(x_T|y_T)$ as discussed next.

6.3.1.3 Kalman-Rauch-Tung-Striebel Smoothing – Inference of $p(x_t|y_{1:T})$

Like the conditional marginal distribution $p(x_t|y_{1:t})$ the conditional marginal $p(x_t|y_{1:T})$ of an LGSSM is a normal distribution and, hence, can be characterized by its first two central moments $\mu_{x_t|y_{1:T}}$ and $\Sigma_{x_t|y_{1:T}}$. In analogy with (6.84) and (6.85), the aim of the following section is to derive the parameters of

$$p(x_t|y_{1:T}) = N(x_t; \mu_{x_t|y_{1:T}}, \Sigma_{x_t|y_{1:T}}) \quad (t = 1, \dots, T) \quad (6.111)$$

in terms of the parameters of the “succeeding distribution”

$$p(x_{t+1}|y_{1:T}) = N(x_{t+1}; \mu_{x_{t+1}|y_{1:T}}, \Sigma_{x_{t+1}|y_{1:T}}) \quad (t = 1, \dots, T-1) \quad (6.112)$$

A backward recursion starting from the final result of the Kalman Filter recursion $p(x_T|y_{1:T})$ may be derived by noting that for the LGSSM the joint distribution over temporally adjacent latent variables conditioned on $y_{1:T}$ may be written as (see below for details)

$$p(x_t, x_{t+1}|y_{1:T}) = \frac{p_\theta(x_{t+1}|x_t)p_\theta(x_t|y_{1:t})p_\theta(x_{t+1}|y_{1:T})}{p_\theta(x_{t+1}|y_{1:t})} \quad (6.113)$$

In the derivation of the KRTS update equations, we follow (Yu, Shenoy, & Sahani, 2004).

◦ **Derivation of equation (6.113)**

We first note that

$$p(x_t, x_{t+1} | y_{1:T}) = p(x_t | x_{t+1}, y_{1:T}) p(x_{t+1} | y_{1:T}) \quad (6.114)$$

We next note that x_t is independent of y_{t+1} given x_{t+1} , i.e. given x_{t+1} , x_t and y_{t+1} are conditionally independent, and one may write

$$p(x_t | x_{t+1}, y_{1:T}) = p(x_t | x_{t+1}, y_{1:t}) \quad (6.115)$$

The conditional independence of x_t and $y_{t+1:T}$ given x_{t+1} may be derived by using the d-separation criterion in the graphical model framework (p. 378 in (Bishop, 2007)): because every path from x_t to $y_{t+1:T}$ is blocked by the node x_{t+1} , i.e. the arrows on the respective paths meet head-to-tail at the node x_{t+1} (see Figure 3A of the tutorial), x_t is conditional independent of $y_{t+1:T}$ given x_{t+1} . In other words

$$x_t \perp y_{k+1} | x_{t+1} \text{ for } k = t, t+1, \dots, T-1 \quad (6.116)$$

We hence have

$$p(x_t, x_{t+1} | y_{1:T}) = p(x_t | x_{t+1}, y_{1:t}) p(x_{t+1} | y_{1:T}) \quad (6.117)$$

Considering the first term on the right-hand side of the above, we have, again with the factorization properties of the LGSSM

$$p(x_t | x_{t+1}, y_{1:t}) = \frac{p(x_t, x_{t+1} | y_{1:t})}{p(x_{t+1} | y_{1:t})} = \frac{p(x_t | y_{1:t}) p(x_{t+1} | x_t, y_{1:t})}{p(x_{t+1} | y_{1:t})} = \frac{p(x_t | y_{1:t}) p(x_{t+1} | x_t)}{p(x_{t+1} | y_{1:t})} \quad (6.118)$$

where the conditional independence of x_{t+1} and $y_{1:t}$ given x_t follows from

$$p(x_{t+1}, y_{1:t} | x_t) = \frac{p(x_t, x_{t+1}, y_{1:t})}{p(x_t)} = \frac{p(x_t) p(x_{t+1} | x_t) p(y_{1:t} | x_t, x_{t+1})}{p(x_t)} = \frac{p(x_t) p(x_{t+1} | x_t) p(y_{1:t} | x_t)}{p(x_t)} = p(x_{t+1} | x_t) p(y_{1:t} | x_t) \quad (6.119)$$

In the penultimate equality of (6.119), the conditional independence of $y_{1:t}$ and x_{t+1} given x_t can again be derived based on the d-separation criterion: all paths from x_{t+1} to $y_{1:t}$ are blocked by the node x_t , or in other words, the respective paths meet tail-to-tail at the node x_t

□

Taking logarithms and substituting for the distributions on the right hand side of (6.113) then results in

$$\begin{aligned} \ln p(x_t, x_{t+1} | y_{1:T}) &= -\frac{1}{2} (x_{t+1} - Ax_t)^T \Sigma_x^{-1} (x_{t+1} - Ax_t) \\ &\quad -\frac{1}{2} (x_t - \mu_{x_t | y_{1:t}})^T (\Sigma_{x_t | y_{1:t}})^{-1} (x_t - \mu_{x_t | y_{1:t}}) \\ &\quad -\frac{1}{2} (x_{t+1} - \mu_{x_{t+1} | y_{1:T}})^T (\Sigma_{x_{t+1} | y_{1:T}})^{-1} (x_{t+1} - \mu_{x_{t+1} | y_{1:T}}) \\ &\quad +\frac{1}{2} (x_{t+1} - \mu_{x_{t+1} | y_{1:t}})^T (\Sigma_{x_{t+1} | y_{1:t}})^{-1} (x_{t+1} - \mu_{x_{t+1} | y_{1:t}}) + C \end{aligned}$$

(6.120)

where $C \in \mathbb{R}$ denotes a constant independent of x_t and x_{t+1} . Rewriting the above in terms of first and second order in x_t and x_{t+1} and comparing these to a standard Gaussian distribution partition allows for inferring the entries of the conditional covariance matrix $\Sigma_{x_t, x_{t+1} | y_{1:T}}$. This conditional covariance matrix partitions according to

$$\Sigma_{x_t, x_{t+1} | y_{1:T}} = \begin{pmatrix} \Sigma_{x_t | y_{1:T}} & \Sigma_{x_t, x_{t+1} | y_{1:T}} \\ \Sigma_{x_{t+1}, x_t | y_{1:T}} & \Sigma_{x_{t+1} | y_{1:T}} \end{pmatrix} = \begin{pmatrix} \mathbb{C}(x_t, x_t | y_{1:T}) & \mathbb{C}(x_t, x_{t+1} | y_{1:T}) \\ \mathbb{C}(x_t, x_{t+1} | y_{1:T}) & \mathbb{C}(x_{t+1}, x_{t+1} | y_{1:T}) \end{pmatrix} \in \mathbb{R}^{2k \times 2k} \quad (6.121)$$

The diagonal entries of $\Sigma_{x_t, x_{t+1} | y_{1:T}}$ in turn allow for inferring the following recursive KRTS update equations (see below for details)

$$\mu_{x_t | y_{1:T}} = \mu_{x_t | y_{1:t}} + \Sigma_{x_t | y_{1:t}} A^T (A \Sigma_{x_t | y_{1:t}} A^T + \Sigma_x)^{-1} (\mu_{x_{t+1} | y_{1:T}} - A \mu_{x_t | y_{1:t}}) \quad (6.122)$$

and

$$\Sigma_{x_t | y_{1:T}} = \Sigma_{x_t | y_{1:t}} + \left(\Sigma_{x_t | y_{1:t}} A^T (A \Sigma_{x_t | y_{1:t}} A^T + \Sigma_x)^{-1} \right) \left(\Sigma_{x_{t+1} | y_{1:T}} - (A \Sigma_{x_t | y_{1:t}} A^T + \Sigma_x) \right) \left(\Sigma_{x_t | y_{1:t}} A^T (A \Sigma_{x_t | y_{1:t}} A^T + \Sigma_x)^{-1} \right)^T \quad (6.123)$$

Notably, these recursive backward expressions for the smoothed expectation $\mu_{x_t | y_{1:T}}$ and covariance parameters $\Sigma_{x_t | y_{1:T}}$ contain only the filtered conditional expectation $\mu_{x_t | y_{1:t}}$, the filtered conditional covariance $\Sigma_{x_t | y_{1:t}}$, the transition matrix A , and the state noise covariance Σ_x . The ensuing algorithm is referred to as KRTS smoother.

◦ Derivations of equations (6.122) and (6.123)

As outlined above, we derive the KRTS update equations by means of the conditional covariance matrix $\Sigma_{x_t, x_{t+1} | y_{1:T}}$ following the approach of (Yu, Shenoy, & Sahani, 2004). This approach capitalizes on the following two matrix identities

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \quad (6.124)$$

and

$$(I - (A + B)^{-1}A)B^{-1} = (A + B)^{-1} \quad (6.125)$$

for appropriately specified matrices A, B, C, U and V . (6.124) is known as the “Woodbury identity” or “Matrix Inversion Lemma”. (6.125) is verified as

$$(A + B)^{-1}(A + B) = I \quad (6.126)$$

$$\Leftrightarrow (A + B)^{-1}A + (A + B)^{-1}B = I$$

$$\Leftrightarrow (A + B)^{-1}B = I - (A + B)^{-1}A$$

$$\Leftrightarrow (A + B)^{-1} = (I - (A + B)^{-1}A)B^{-1}$$

$$\Leftrightarrow (I - (A + B)^{-1}A)B^{-1} = (A + B)^{-1}$$

Further, the approach of (Yu, Shenoy, & Sahani, 2004) uses the formulation of the quadratic form in the exponent of multivariate Gaussian distributions as introduced in Supplement Section 2 and written here for

$$p(z_1, z_2) = N\left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}; \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Lambda^{-1}\right) \quad (6.127)$$

as

$$\begin{aligned} \ln p(z_1, z_2) &= -\frac{1}{2} \begin{pmatrix} z_1 - \mu_1 \\ z_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \begin{pmatrix} z_1 - \mu_1 \\ z_2 - \mu_2 \end{pmatrix} + C \\ &= -\frac{1}{2} z_1^T \Lambda_{11} z_1 - \frac{1}{2} z_1^T \Lambda_{12} z_2 - \frac{1}{2} z_2^T \Lambda_{21} z_1 - \frac{1}{2} z_2^T \Lambda_{22} z_2 + z_2^T (\Lambda_{21} \mu_1 + \Lambda_{22} \mu_2) + z_1^T (\Lambda_{11} \mu_1 + \Lambda_{12} \mu_2) + C \end{aligned} \quad (6.128)$$

with suitably chosen vectors z_1, z_2, μ_1, μ_2 and block matrix Λ , denoting the inverse of the covariance matrix of $p(z_1, z_2)$ and $C \in \mathbb{R}$. In contrast to Supplement Section 2 we here use numerical indices to avoid confusion with the nomenclature for latent and observed variables in the LGSSM context. In addition (Yu, Shenoy, & Sahani, 2004) note that for the inversion of block matrices, the following equalities hold. Let

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} := \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1} \quad (6.129)$$

and

$$\Gamma_{11} := \Lambda_{11} - \Lambda_{12} \Lambda_{22}^{-1} \Lambda_{21} \quad \text{and} \quad \Gamma_{22} := \Lambda_{22} - \Lambda_{21} \Lambda_{11}^{-1} \Lambda_{12} \quad (6.130)$$

Then

$$\begin{aligned} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1} &= \begin{pmatrix} \Gamma_{11}^{-1} & -\Gamma_{11}^{-1} \Lambda_{12} \Lambda_{22}^{-1} \\ -\Lambda_{22}^{-1} \Lambda_{21} \Gamma_{11}^{-1} & \Gamma_{22}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \Lambda_{11}^{-1} + \Lambda_{11}^{-1} \Lambda_{12} \Gamma_{22}^{-1} \Lambda_{21} \Lambda_{11}^{-1} & -\Gamma_{11}^{-1} \Lambda_{12} \Lambda_{22}^{-1} \\ -\Lambda_{22}^{-1} \Lambda_{21} \Gamma_{11}^{-1} & \Lambda_{22}^{-1} + \Lambda_{22}^{-1} \Lambda_{21} \Gamma_{11}^{-1} \Lambda_{12} \Lambda_{22}^{-1} \end{pmatrix} \end{aligned} \quad (6.131)$$

which may be shown using standard matrix algebra.

Based on the set of preliminaries (6.124) – (6.131), equations (6.122) and (6.123) may now be derived from (6.120) as follows. As a first step, the right hand side of (6.120) is reordered in terms of second-order terms in x_{t+1} , first-order mixed terms in x_{t+1} and x_t , second-order terms in x_t and first-order terms in x_t . This reordering results in

$$\begin{aligned} \ln p(x_t, x_{t+1} | y_{1:T}) &= -\frac{1}{2} x_{t+1}^T \left(\Sigma_x^{-1} - (\Sigma_{x_{t+1}|y_{1:t}})^{-1} + (\Sigma_{x_{t+1}|y_{1:T}})^{-1} \right) x_{t+1} \\ &\quad -\frac{1}{2} x_{t+1}^T (-\Sigma_x^{-1} A) x_t - \frac{1}{2} x_t^T (-A^T \Sigma_x^{-1}) x_{t+1} \\ &\quad -\frac{1}{2} x_t^T \left(A^T \Sigma_x^{-1} A + (\Sigma_{x_t|y_{1:t}})^{-1} \right) x_t \\ &\quad + x_t^T \left((\Sigma_{x_t|y_{1:t}})^{-1} \mu_{x_t|y_{1:t}} \right) + C \end{aligned} \quad (6.132)$$

Term-by-term comparison of (6.132) and (6.128) then allows for inferring the entries of the inverse covariance matrix of $p(x_{t+1}, x_t | y_{1:T})$ to be given as

$$p(x_{t+1}, x_t | y_{1:T}) = \begin{pmatrix} \Sigma_{x_{t+1}|y_{1:T}} & \Sigma_{x_{t+1},t|y_{1:T}} \\ \Sigma_{x_t,t+1|y_{1:T}} & \Sigma_{x_t|y_{1:T}} \end{pmatrix} = \begin{pmatrix} \Sigma_x^{-1} - \Sigma_{x_{t+1}|y_{1:t}}^{-1} + \Sigma_{x_{t+1}|y_{1:T}}^{-1} & -\Sigma_x^{-1}A \\ -A^T\Sigma_x^{-1} & A^T\Sigma_x^{-1}A + \Sigma_{x_t|y_{1:t}}^{-1} \end{pmatrix}^{-1} \quad (6.133)$$

To obtain the covariance matrix of $p(x_{t+1}, x_t | y_{1:T})$, the block matrix on the right-hand side of (6.133) has to be inverted. To achieve this, and simultaneously derive (6.123) (Yu, Shenoy, & Sahani, 2004), first simplify Λ_{22}^{-1} and $\Lambda_{22}^{-1}\Lambda_{21}$ in (6.131) in terms of the entries in (6.133). Specifically, using the matrix inversion lemma and (2.10)

$$\begin{aligned} \Lambda_{22}^{-1} &= (A^T\Sigma_x^{-1}A + \Sigma_{x_t|y_{1:t}}^{-1})^{-1} = \Sigma_{x_t|y_{1:t}} - \Sigma_{x_t|y_{1:t}}A^T(\Sigma_x + A\Sigma_{x_t|y_{1:T}}A^T)^{-1}A\Sigma_{x_t|y_{1:t}} \\ &= \Sigma_{x_t|y_{1:t}} - \Sigma_{x_t|y_{1:t}}A^T(\Sigma_{x_{t+1}|y_{1:t}})^{-1}A\Sigma_{x_t|y_{1:t}} \end{aligned} \quad (6.134)$$

and defining

$$J_t := \Sigma_{x_t|y_{1:t}}A^T(\Sigma_{x_{t+1}|y_{1:t}})^{-1} \quad (6.135)$$

we obtain

$$\Lambda_{22}^{-1} = \Sigma_{x_t|y_{1:t}} - J_t\Sigma_{x_{t+1}|y_{1:t}}J_t^T \quad (6.136)$$

Likewise, by means of the matrix identity

$$(I - (A + B)^{-1}A)B^{-1} = (A + B)^{-1} \quad (6.137)$$

we obtain

$$\begin{aligned} \Lambda_{22}^{-1}\Lambda_{21} &= -(\Sigma_{x_t|y_{1:t}} - J_t\Sigma_{x_{t+1}|y_{1:t}}J_t^T)A^T\Sigma_x^{-1} \\ &= -\Sigma_{x_t|y_{1:t}}A^T(I - (\Sigma_x + A\Sigma_{x_t|y_{1:T}}A^T)^{-1}A\Sigma_{x_t|y_{1:t}}A^T)\Sigma_x^{-1} \\ &= -\Sigma_{x_t|y_{1:t}}A^T(\Sigma_x + A\Sigma_{x_t|y_{1:t}}A^T)^{-1} \\ &= -J_t \end{aligned} \quad (6.138)$$

After this simplification, the block-matrix on the right-hand side of (6.133) may be inverted to obtain the entries of the covariance matrix of $p(x_{t+1}, x_t | y_{1:T})$ according to (6.131). Specifically, we have

$$\Sigma_{x_{t+1}|y_{1:T}} = \Gamma_{11}^{-1} \quad (6.139)$$

and therefore

$$\begin{aligned} \Sigma_{x_t|y_{1:T}} &:= \Lambda_{22}^{-1} + \Lambda_{22}^{-1}\Lambda_{21}\Gamma_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1} \\ &= (\Sigma_{x_t|y_{1:t}} - J_t\Sigma_{x_{t+1}|y_{1:t}}J_t^T) + (-J_t)\Sigma_{x_{t+1}|y_{1:T}}(-J_t^T) \\ &= \Sigma_{x_t|y_{1:t}} + J_t(\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}})J_t^T \end{aligned} \quad (6.140)$$

$$= \Sigma_{x_t|y_{1:t}} + \left(\Sigma_{x_t|y_{1:t}} A^T (\Sigma_{x_{t+1}|y_{1:t}})^{-1} \right) (\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}}) \left(\Sigma_{x_t|y_{1:t}} A^T (\Sigma_{x_{t+1}|y_{1:t}})^{-1} \right)^T$$

Noting that with the Gaussian linear transformation theorem

$$\Sigma_{x_t|y_{1:t-1}} = A \Sigma_{x_{t-1}|y_{1:t-1}} A^T + \Sigma_x \quad (6.141)$$

and thus

$$\Sigma_{x_{t+1}|y_{1:t}} = A \Sigma_{x_t|y_{1:t}} A^T + \Sigma_x \quad (6.142)$$

we thus have (6.123), i.e. a backward recursive expression for $\Sigma_{x_t|y_{1:T}}$ in terms of the filtered covariance matrices $\Sigma_{x_t|y_{1:t}}$ and the parameters of the LGSSM

$$\Sigma_{x_t|y_{1:T}} = \Sigma_{x_t|y_{1:t}} + \left(\Sigma_{x_t|y_{1:t}} A^T (A \Sigma_{x_t|y_{1:t}} A^T + \Sigma_x)^{-1} \right) (\Sigma_{x_{t+1}|y_{1:T}} - (A \Sigma_{x_t|y_{1:t}} A^T + \Sigma_x)) \left(\Sigma_{x_t|y_{1:t}} A^T (A \Sigma_{x_t|y_{1:t}} A^T + \Sigma_x)^{-1} \right)^T \quad (6.143)$$

To find the expectation parameter, we compare the last terms in (6.128) and (6.132), i.e.

$$z_2^T (\Lambda_{21} \mu_1 + \Lambda_{22} \mu_2) \quad \text{and} \quad x_t^T (\Sigma_{x_t|y_{1:t}}^{-1} \mu_{x_t|y_{1:t}}) \quad (6.144)$$

Apparently, we have

$$\mu_1 := \mu_{x_{t+1}|y_{1:T}} \quad \text{and} \quad \mu_2 := \mu_{x_t|y_{1:T}} \quad (6.145)$$

We hence infer

$$\Lambda_{21} \mu_{x_{t+1}|y_{1:T}} + \Lambda_{22} \mu_{x_t|y_{1:T}} = \Sigma_{x_t|y_{1:t}}^{-1} \mu_{x_t|y_{1:t}} \quad (6.146)$$

Solving for $\mu_{x_t|y_{1:T}}$, we obtain

$$\mu_{x_t|y_{1:T}} = -\Lambda_{22}^{-1} \Lambda_{21} \mu_{x_{t+1}|y_{1:T}} + \Lambda_{22}^{-1} \Sigma_{x_t|y_{1:t}}^{-1} \mu_{x_t|y_{1:t}} \quad (6.147)$$

Next, using

$$\Lambda_{22}^{-1} \Lambda_{21} = -J_t \quad \text{and} \quad \Lambda_{22}^{-1} = \Sigma_{x_t|y_{1:t}} - J_t \Sigma_{x_{t+1}|y_{1:t}} J_t^T \quad (6.148)$$

we obtain

$$\mu_{x_t|y_{1:T}} = J_t \mu_{x_{t+1}|y_{1:T}} + (I - J_t A) \mu_{x_t|y_{1:t}} \quad (6.149)$$

$$= J_t \mu_{x_{t+1}|y_{1:T}} + \mu_{x_t|y_{1:t}} - J_t A \mu_{x_t|y_{1:t}}$$

$$= \mu_{x_t|y_{1:t}} + J_t (\mu_{x_{t+1}|y_{1:T}} - A \mu_{x_t|y_{1:t}})$$

$$= \mu_{x_t|y_{1:t}} + (\Sigma_{x_t|y_{1:t}} A^T (\Sigma_{x_{t+1}|y_{1:t}})^{-1}) (\mu_{x_{t+1}|y_{1:T}} - A \mu_{x_t|y_{1:t}})$$

$$= \mu_{x_t|y_{1:t}} + (\Sigma_{x_t|y_{1:t}} A^T (A \Sigma_{x_t|y_{1:t}} A^T + \Sigma_x)^{-1}) (\mu_{x_{t+1}|y_{1:T}} - A \mu_{x_t|y_{1:t}})$$

□

As for the Kalman filter equations, (6.122) and (6.123) may be written more succinctly as

$$\mu_{x_t|y_{1:T}} = \mu_{x_t|y_{1:t}} + J_t(\mu_{x_{t+1}|y_{1:T}} - A\mu_{x_t|y_{1:t}}) \quad (6.150)$$

and

$$\begin{aligned} \Sigma_{x_t|y_{1:T}} &= \Sigma_{x_t|y_{1:t}} + J_t \left(\Sigma_{x_{t+1}|y_{1:T}} - (A\Sigma_{x_t|y_{1:t}}A^T + \Sigma_x) \right) J_t^T \\ &= \Sigma_{x_t|y_{1:t}} + J_t (\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}}) J_t^T \end{aligned} \quad (6.151)$$

using the definition

$$J_t := \Sigma_{x_t|y_{1:t}} A^T (A\Sigma_{x_t|y_{1:t}}A^T + \Sigma_x)^{-1} \quad (6.152)$$

The recursion is initialized using the Kalman filtering result for $\mu_{x_T|y_{1:T}}$ and $\Sigma_{x_T|y_{1:T}}$.

6.3.1.4 Inferring $p(x_{t-1}, x_t | y_{1:T})$

The expectation parameter components of $p(x_{t-1}, x_t | y_{1:T})$ correspond to the expectation parameters of $p(x_{t-1} | y_{1:T})$ and $p(x_t | y_{1:T})$, respectively, i.e., we have

$$\mu_{x_{t-1:t}|y_{1:T}} = (\mu_{x_{t-1}|y_{1:T}}, \mu_{x_t|y_{1:T}})^T \quad (6.153)$$

where the right-hand side components are evaluated based on the KRTS smoothing equations as discussed in section 6.3.1.3. For the covariance matrix parameter of $p(x_{t-1}, x_t | y_{1:T})$, we note that according to (6.131), we have

$$\Sigma_{x_{t,t+1}|y_{1:T}} = -\Lambda_{22}^{-1} \Lambda_{21} \Gamma_{11}^{-1} \quad (6.154)$$

and with

$$\Sigma_{x_{t+1}|y_{1:T}} = \Gamma_{11}^{-1} \text{ and } \Lambda_{22}^{-1} \Lambda_{21} = -J_t \quad (6.155)$$

We thus have the following equivalences for the covariance matrix of $p(x_{t-1}, x_t | y_{1:T})$:

$$\begin{aligned} \Sigma_{x_{t,t+1}|y_{1:T}} &= \Sigma_{x_{t+1}|y_{1:T}} J_t^T \\ \Leftrightarrow \Sigma_{x_{t-1:t}|y_{1:T}} &= \Sigma_{x_t|y_{1:T}} J_{t-1}^T \\ \Leftrightarrow \left(\Sigma_{x_{t-1:t}|y_{1:T}} \right)^T &= \left(\Sigma_{x_t|y_{1:T}} J_{t-1}^T \right)^T \\ \Leftrightarrow \Sigma_{x_{t-1:t}|y_{1:T}} &= J_{t-1} \Sigma_{x_t|y_{1:T}} \end{aligned} \quad (6.156)$$

where the last line corresponds to the desired recursion for $\Sigma_{x_{t-1:t}|y_{1:T}}$, which is initialized as

$$\Sigma_{x_{T-1:T}|y_{1:T}} = J_{T-1} \Sigma_{x_T|y_{1:T}} \quad (6.157)$$

6.3.1.5 Summary and expectations

The Kalman filter and KRTS smoothing recursions derived in Sections 6.3.1.1 – 6.3.1.4 are shown in pseudocode form in Table 4. Simplification of these equations for the univariate case $x_{1:T}, y_{1:T} \in \mathbb{R}$ $A := a \in \mathbb{R}$, $B := b \in \mathbb{R}$, $\Sigma_x := \sigma_x > 0$ and $\Sigma_y := \sigma_y$ then yield equations (32) – (38) of the tutorial:

$$\mu_{x_t|y_{1:t}} := A\mu_{x_{t-1}|y_{1:t-1}} + K(y_t - BA\mu_{x_{t-1}|y_{1:t-1}}) \Leftrightarrow \mu_{x_t|y_{1:t}} := a\mu_{x_{t-1}|y_{1:t-1}} + k_t(y_t - ba\mu_{x_{t-1}|y_{1:t-1}}) \quad (6.158)$$

$$\Sigma_{x_t|y_{1:t}} := (I - K_t B) P_t \Leftrightarrow \sigma_{x_t|y_{1:t}}^2 := (1 - k_t b)(a^2 \sigma_{x_{t-1}|y_{1:t-1}}^2 + \sigma_x^2) \quad (6.159)$$

$$\mu_{x_t|y_{1:T}} := \mu_{x_t|y_{1:t}} + J_t(\mu_{x_{t+1}|y_{1:T}} - A\mu_{x_t|y_{1:t}}) \Leftrightarrow \mu_{x_t|y_{1:T}} := \mu_{x_t|y_{1:t}} + j_t(\mu_{x_{t+1}|y_{1:T}} - a\mu_{x_t|y_{1:t}}) \quad (6.160)$$

$$\Sigma_{x_t|y_{1:T}} := \Sigma_{x_t|y_{1:t}} + J_t(\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}})J_t^T \Leftrightarrow \sigma_{x_t|y_{1:T}}^2 := \sigma_{x_t|y_{1:t}}^2 + j_t^2(\sigma_{x_{t+1}|y_{1:T}}^2 - \sigma_{x_{t+1}|y_{1:t}}^2) \quad (6.161)$$

$$\Sigma_{x_{t-1,t}|y_{1:T}} := J_{t-1}\Sigma_{x_t|y_{1:T}} \Leftrightarrow \sigma_{x_{t-1,t}|y_{1:T}}^2 := j_{t-1}\sigma_{x_t|y_{1:T}}^2 \quad (6.162)$$

The variational update equations derived in Section 6.2 capitalize on the availability of the expectations of x_t and x_t^2 under the marginal variational distributions $q^{(i)}(x_t)$ and the expectation of $x_{t-1}x_t$ under the variational joint marginal distributions $q^{(i)}(x_{t-1:t})$ ($t = 2, \dots, T$). As will be seen in Section 6.3.2, a necessary prerequisite for their evaluation is the evaluation of these expectations under the conditional LGSSM distribution $p(x_{1:T}|y_{1:T})$ for known parameter θ . For the multivariate case, these evaluate in terms of the parameters of $p(x_t|y_{1:T})$ and $p(x_{t-1}, x_t|y_{1:T})$ for $t = 1, \dots, T$ to

$$\langle x_t \rangle_{p(x_t|y_{1:T})} = \mu_{x_t|y_{1:T}} \quad (6.163)$$

$$\langle x_t x_t^T \rangle_{p(x_t|y_{1:T})} = \mu_{x_t|y_{1:T}} \mu_{x_t|y_{1:T}}^T + \Sigma_{x_t|y_{1:T}} \quad (6.164)$$

and

$$\langle x_{t-1} x_t^T \rangle_{p(x_{t-1}, x_t|y_{1:T})} = \mu_{x_{t-1}|y_{1:T}} \mu_{x_t|y_{1:T}}^T + \Sigma_{x_{t-1,t}|y_{1:T}} \quad (6.165)$$

◦ Derivation of euqations (6.163) – (6.164)

The equations follow directly from the well-known identity

$$\mathbb{C}(x, y) = \mathbb{E}(xy^T) - \mathbb{E}(x)\mathbb{E}(y^T) \Rightarrow \mathbb{E}(xx^T) = \mathbb{C}(x, x) + \mathbb{E}(x)\mathbb{E}(x^T)$$

(see Supplement equation (1.21) and Bishop (2008) p. 20).

□

function Kalman Filter (Input, Output)

Input $y_{1:T}, \mu_1, \Sigma_1, A, \Sigma_x, B, \Sigma_y$
 Output $\mu_{x_t|y_{1:t}}$ and $\Sigma_{x_t|y_{1:t}}$ for $t = 1, \dots, T$

% Initialization = Bayesian linear model inference for $t = 1$

$$\mu_{x_1|y_1} := \mu_{x_1} + \Sigma_{x_1} B^T (B \Sigma_{x_1} B^T + \Sigma_y)^{-1} (y_1 - B \mu_{x_1})$$

$$\Sigma_{x_1|y_1} := \Sigma_{x_1} - \Sigma_{x_1} B^T (B \Sigma_{x_1} B^T + \Sigma_y)^{-1} B \Sigma_{x_1}$$

% Recursion

for $t = 2, \dots, T$ set

$$P_t := A \Sigma_{x_{t-1}|y_{1:t-1}} A^T + \Sigma_x$$

$$K_t := P_t B^T (B P_t B^T + \Sigma_y)^{-1}$$

$$\mu_{x_t|y_{1:t}} := A \mu_{x_{t-1}|y_{1:t-1}} + K_t (y_t - B A \mu_{x_{t-1}|y_{1:t-1}})$$

$$\Sigma_{x_t|y_{1:t}} := (I - K_t B) P_t (I - K_t B)^T + K_t \Sigma_y K_t^T$$

function KRTS Smoother (Input, Output)

Input $y_{1:T}, A, \Sigma_x$ and $\mu_{x_t|y_{1:t}}, \Sigma_{x_t|y_{1:t}}$ for $t = 1, \dots, T$
 Output $\mu_{x_t|y_{1:T}}, \Sigma_{x_t|y_{1:T}}$ for $t = 1, \dots, T$ and $\Sigma_{x_{t-1:t}|y_{1:T}}$ for $t = 2, \dots, T$

% Initialization = Kalman filter results for $t = T$

$$\mu_{x_T|y_{1:T}} := \mu_{x_T|y_{1:T}}$$

$$\Sigma_{x_T|y_{1:T}} := \Sigma_{x_T|y_{1:T}}$$

% Recursion

for $t = T - 1, T - 2, \dots, 1$ set

$$J_t := \Sigma_{x_t|y_{1:t}} A^T (A \Sigma_{x_t|y_{1:t}} A^T + \Sigma_x)^{-1}$$

$$\mu_{x_t|y_{1:T}} := \mu_{x_t|y_{1:t}} + J_t (\mu_{x_{t+1}|y_{1:T}} - A \mu_{x_t|y_{1:t}})$$

$$\Sigma_{x_t|y_{1:T}} := \Sigma_{x_t|y_{1:t}} + J_t (\Sigma_{x_{t+1}|y_{1:T}} - \Sigma_{x_{t+1}|y_{1:t}}) J_t^T$$

$$\Sigma_{x_{t-1:t}|y_{1:T}} := J_{t-1} \Sigma_{x_t|y_{1:T}}$$

Table 5. Pseudocode for the Kalman filter and KRTS smoother recursions.

6.3.1.6 Matlab implementation

Below follows the Matlab implementation of the pseudocode shown in Table 5 as used in the tutorial.

```
function [mu_f, Sigma_f] = lgssm_filter(y,mu_x_1,Sigma_x_1,A,B,Sigma_x,Sigma_y)

% This function implements the forward pass for inference in a linear
% Gaussian state space model, also known as the Kalman Filter. It
% capitalizes on the expressions for the conditional expectation and
% covariance as discussed in section 5 of the tutorial. Specifically, for
% the covariance update, Joseph's symmetrized form is employed.
%
% Inputs
%       y           : p x T data array
%       mu_x_1      : k x 1 initial latent variable expectation
%       Sigma_x_1    : k x k initial latent variable covariance
%       A           : k x k transition matrix
%       B           : p x k emission matrix
%       Sigma_x      : k x k latent variable noise covariance
%       Sigma_y      : p x p latent variable noise covariance
%
% Outputs
%       mu_f        : k x T      filtered mean of p(x_t|y_(1:t))
%       Sigma_f      : k x k x T filtered covariance of p(x_t|y_(1:t))
%
% Copyright (C) Dirk Ostwald
% -----
% recover system and data dimensionalities
k      = size(A,1);
p      = size(y,1);
T      = size(y,2);

% initialize conditional expectation and covariance arrays
mu_f   = NaN(k,T);
Sigma_f = NaN(k,k,T);

% Initialization
% -----
mu_f(:,1) = mu_x_1 + Sigma_x_1*B'*inv(B*Sigma_x_1*B' + Sigma_y)*(y(:,1) - B*mu_x_1);
Sigma_f(:, :, 1) = Sigma_x_1 - Sigma_x_1*B'*inv(B*Sigma_x_1*B' + Sigma_y)*B*Sigma_x_1;

% filtered covariance symmetry check
if ~isequal(Sigma_f(:, :, 1), Sigma_f(:, :, 1)')
    Sigma_f(:, :, 1) = 1/2*(Sigma_f(:, :, 1) + Sigma_f(:, :, 1)');
end

% Recursion
% -----

% cycle over remaining time-points
for t = 2:T

    % predictor and Kalman gain matrices
    P      = A*Sigma_f(:, :, t-1)*A' + Sigma_x;
    K      = P*B'*inv(B*P*B' + Sigma_y);

    % filtered expectation and covariance updates, Joseph's form
    mu_f(:, t) = A*mu_f(:, t-1) + K*(y(:, t) - B*A*mu_f(:, t-1));
    Sigma_f(:, :, t) = (eye(k) - K*B)*P*(eye(k) - K*B)' + K*Sigma_y*K';

    % filtered covariance symmetry check
    if ~isequal(Sigma_f(:, :, t), Sigma_f(:, :, t)')
        Sigma_f(:, :, t) = 1/2*(Sigma_f(:, :, t) + Sigma_f(:, :, t)');
    end
end
end
```

```

function [mu_s, Sigma_s, Sigma_s_c] = lgssm_smooth(y, mu_f, A, Sigma_f, Sigma_x)

% This function implements the backward pass for inference in a linear
% Gaussian state space model, also known as the Kalman-Rauch-Tung-Striebel
% smoother, and here in addition returning the cross-moment covariance of
%  $p(x_t, x_{t+1} | y_{1:T})$ 
%
% Inputs
%
%     y           : p x T data array
%     mu_f        : k x T filtered latent variable expectation
%     A           : k x k transition matrix
%     Sigma_f     : k x k filtered latent variable covariance
%     Sigma_x     : k x k latent variable noise covariance
%
% Outputs
%
%     mu_s        : k x 1 smoothed mean of  $p(x_t | y_{1:T})$ 
%     Sigma_s     : k x k x T filtered covariance of  $p(x_t | y_{1:T})$ 
%     Sigma_s_c   : k x k x T-1 filtered cross moment covariance of
%                   $p(x_t, x_{t+1} | y_{1:T})$  for  $t = 1, \dots, T$ 
%
% Copyright (C) Dirk Ostwald
% -----
% recover system and data dimensionalities
k      = size(A,1);
p      = size(y,1);
T      = size(y,2);

% initialize conditional expectation, covariance and cross moment
% covariance arrays
mu_s   = NaN(k,T);
Sigma_s = NaN(k,k,T);
Sigma_s_c = NaN(k,k,T-1);

% filter initialization
% -----
mu_s(:,T)      = mu_f(:,T);
Sigma_s(:,:,T) = Sigma_f(:,:,T);

% filter recursion
% -----
for t = T-1:-1:1

    % smoothing matrix J
    J = Sigma_f(:,:,t)*A'*inv(A*Sigma_f(:,:,t)*A' + Sigma_x);

    % smoothed expectation and covariance updates
    mu_s(:,t)      = mu_f(:,t) + J*(mu_s(:,t+1) - A*mu_f(:,t));
    Sigma_s(:,:,t) = Sigma_f(:,:,t) + J*(Sigma_s(:,:,t+1) - Sigma_f(:,:,t+1))*J';

    % smoothed covariance symmetry check
    if ~isequal(Sigma_s(:,:,t),Sigma_s(:,:,t)')
        Sigma_s(:,:,t) = 1/2*(Sigma_s(:,:,t) + Sigma_s(:,:,t)');
    end

    % cross moment
    Sigma_s_c(:,:,t) = J*Sigma_s(:,:,t+1);

end
end

```

6.3.2 Unified inference

In this section, we consider the variational Bayes update equation

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(y_{1:T}, x_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right) \quad (6.166)$$

for the general case that θ is an unobserved random variable and thus the expectation over $q^{(i+1)}(\theta)$ and the marginal distribution $p(\theta)$ in the argument of the exponential above cannot be neglected. We first note that from (6.71), (6.166) is equivalent to

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(x_{1:T}, y_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)}\right) \quad (6.167)$$

which from (6.73), in turn, is equivalent to

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln p(x_{1:T} | y_{1:T}, \theta) \rangle_{q^{(i+1)}(\theta)}\right) \quad (6.168)$$

Further, the argument of the logarithm above can be re-parameterized according to (6.74) as

$$q^{(i+1)}(x_{1:T}) \propto \exp\left(\langle \ln \left(p(x_1 | y_{1:T}) \prod_{t=2}^T \frac{p(x_{t-1}, x_t | y_{1:T})}{p(x_{t-1} | y_{1:T})} \right) \rangle_{q^{(i+1)}(\theta)}\right) \quad (6.169)$$

which corresponds to the parameterization exploited by the KRTS inference machinery. While the proportionality statements of (6.166) – (6.169) are equivalent, following (D. Barber & Chiappa, 2007), we focus on the formulation of the variational update rule in the form of (6.167) below, sometimes referred to as the “expected energy formulation”. For the purposes of inferring the parameters of the right-hand side of (6.167) using the KRTS inference framework, one may think that, for the current purposes of random θ the parameter expectations under the variational distributions, or in other words, $\langle \theta \rangle_{q^{(i+1)}(\theta)}$, may be forwarded to the KRTS smoothing algorithm. However, this is not appropriate, as one can show, that for the LGSSM the expectation of the energy function under the current variational distribution over parameters $q^{(i+1)}(\theta)$ does not equal the energy function with averaged parameters. In other words, in general,

$$\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q^{(i+1)}(\theta)} \neq \ln p(y_{1:T}, x_{1:T} | \langle \theta \rangle_{q^{(i+1)}(\theta)}) \quad (6.170)$$

Barber and Chiappa (D. Barber & Chiappa, 2007) have shown how, nevertheless, standard inference algorithms may be applied for the VB update of $q^{(i+1)}(x_{1:T})$. In the following, we further discuss the inequality (6.170) as the “mean and fluctuation decomposition theorem” in Section 6.3.2.1 and the solution provided by (D. Barber & Chiappa, 2007) in form of the “unified inference theorem” in Section 6.3.2.2. As in Section 6.3.1, we use the general multivariate LGSSM form (cf. equation 6.77) throughout and only derive the special univariate LGSSM case for the tutorial example at the end of this section. In contrast to Section 6.3.1, however, we here use a precision matrix rather than covariance matrix parameterization of the Gaussian distributions involved, because we posited precision parameter marginal (prior) distributions in (6.5) and (6.7) and now are required to take these explicitly into account. With $\theta := \{\mu_1, \Lambda_1, A, \Lambda_x, B, \Lambda_y\}$ we thus reexpress equation (6.77) for the current section as

$$p_\theta(x_{1:T}, y_{1:T}) = p_{\mu_1, \Lambda_1}(x_1) \prod_{t=2}^T p_{A, \Lambda_x}(x_t | x_{t-1}) \prod_{t=1}^T p_{B, \Lambda_y}(y_t | x_t) \quad (6.171)$$

where

$$p(x_1) = N(x_1; \mu, \Lambda_1^{-1}), p(x_t | x_{t-1}) = N(x_t; A x_{t-1}, \Lambda_x^{-1}) \text{ and } p(y_t | x_t) = N(y_t; B x_t, \Lambda_y^{-1}) \quad (6.172)$$

6.3.2.1 The mean and fluctuation decomposition theorem

The inequality (6.170) is summarized in (D. Barber & Chiappa, 2007) as the “mean and fluctuation decomposition theorem.” Specifically, it can be shown that the expectation of the conditional energy function under $q^{(i+1)}(\theta)$ can be written as the sum of two components: (1) the joint distribution of $x_{1:T}$ and $y_{1:T}$ conditioned on the parameter expectations $\langle \theta \rangle_{q^{(i+1)}(\theta)}$ and (2) “fluctuation” terms. For ease of notation, we omit the superscript $(i+1)$ for the variational distributions over the elements of θ in the formulations below. We have the following theorem:

Mean and fluctuation decomposition theorem

The following decomposition holds

$$\langle \ln p(x_{1:T}, y_{1:T} | \theta) \rangle_{q(\theta)} = \ln p(x_{1:T}, y_{1:T} | \langle \theta \rangle_{q(\theta)}) + F_{A, \Lambda_x} + F_{B, \Lambda_y} \quad (6.171)$$

where, with a constant $C \in \mathbb{R}$

$$\begin{aligned} \ln p(x_{1:T}, y_{1:T} | \langle \theta \rangle_{q(\theta)}) &= -\frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \\ &\quad - \frac{1}{2} \sum_{t=2}^T (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_x)} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \\ &\quad - \frac{1}{2} \sum_{t=1}^T (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t)^T \langle \Lambda_y \rangle_{q(\Lambda_y)} (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t) \\ &\quad + C \end{aligned} \quad (6.172)$$

and

$$F_{A, \Lambda_x} := -\frac{1}{2} \sum_{t=1}^{T-1} x_t^T (\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)}) x_t \quad (6.174)$$

$$F_{B, \Lambda_y} := -\frac{1}{2} \sum_{t=1}^T x_t^T (\langle B^T \Lambda_y B \rangle_{q(B, \Lambda_y)} - \langle B^T \rangle_{q(B|\Lambda_y)} \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)}) x_t \quad (6.175)$$

□

Note that the decomposition implies the inequality of (6.170) for non-zero terms of F_{A, Λ_x} and F_{B, Λ_y} . Given that the “fluctuation” terms F_{A, Λ_x} and F_{B, Λ_y} do not vanish except under the assumption of infinite precision of the variational distribution (i.e. the scenario discussed in Section 6.3.1), standard LGSSM inference algorithms supplied with the expected LGSSM parameters $\langle \theta \rangle_{q^{(i+1)}(\theta)}$ are thus not appropriate for inference of the variational update $q^{(i+1)}(x_{1:T})$.

◦ Proof of the mean and fluctuation decomposition theorem

We start by reformulating the left-hand side of (6.171) based on the conditional independence properties of the LGSSM.

$$\begin{aligned} \langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} &= \langle \ln p(x_{1:T}, y_{1:T} | A, \Lambda_x, B, \Lambda_y) \rangle_{q(A, \Lambda_x, B, \Lambda_y)} \\ &= \langle \ln p(x_{1:T}, y_{1:T} | A, \Lambda_x, B, \Lambda_y) \rangle_{q(A, \Lambda_x) q(B, \Lambda_y)} \\ &= \langle \ln(p(x_1) \prod_{t=2}^T p(x_t | x_{t-1}, A, \Lambda_x) \prod_{t=1}^T p(y_t | x_t, B, \Lambda_y)) \rangle_{q(A, \Lambda_x) q(B, \Lambda_y)} \end{aligned}$$

(6.176)

where the factorization of $q(A, \Lambda_x, B, \Lambda_y)$ follows from the properties of the LGSSM (Beal, 2003). Substitution of the respective probability density functions in (6.176) then yields

$$\begin{aligned}
\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} &= \langle \ln \left(N(x_1; \mu_1, \Lambda_1^{-1}) \prod_{t=2}^T N(x_t; Ax_{t-1}, \Lambda_x^{-1}) \prod_{t=1}^T N(y_t; Bx_t, \Lambda_y^{-1}) \right) \rangle_{q(A, \Lambda_x)q(B, \Lambda_y)} \\
&= \langle \ln \left((2\pi)^{-\frac{k}{2}} |\Lambda_1^{-1}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \right) \right) \rangle_{q(A, \Lambda_x)q(B, \Lambda_y)} \\
&\quad + \langle \ln \left(\prod_{t=2}^T (2\pi)^{-\frac{k}{2}} |\Lambda_x^{-1}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \right) \right) \rangle_{q(A, \Lambda_x)q(B, \Lambda_y)} \\
&\quad + \langle \ln \left(\prod_{t=1}^T (2\pi)^{-\frac{p}{2}} |\Lambda_y^{-1}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (y_t - Bx_t)^T \Lambda_y (y_t - Bx_t) \right) \right) \rangle_{q(A, \Lambda_x)q(B, \Lambda_y)}
\end{aligned} \tag{6.177}$$

Evaluation of the expectations with respect to the variational distributions whose arguments do not feature in the respective terms then yields

$$\begin{aligned}
\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} &= \ln \left((2\pi)^{-\frac{k}{2}} |\Lambda_1^{-1}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \right) \right) \\
&\quad + \langle \ln \left(\prod_{t=2}^T (2\pi)^{-\frac{k}{2}} |\Sigma_x|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \right) \right) \rangle_{q(A, \Lambda_x)} \\
&\quad + \langle \ln \left(\prod_{t=1}^T (2\pi)^{-\frac{p}{2}} |\Lambda_y^{-1}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (y_t - Bx_t)^T \Lambda_y (y_t - Bx_t) \right) \right) \rangle_{q(B, \Lambda_y)}
\end{aligned} \tag{6.178}$$

Evaluating the remaining expectations and the logarithmic terms then yields

$$\begin{aligned}
\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} &= -\frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\Lambda_1^{-1}| - \frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \\
&\quad + \left\langle -\frac{k(T-1)}{2} \ln 2\pi - \frac{T-1}{2} \ln |\Lambda_x^{-1}| - \frac{1}{2} \sum_{t=2}^T (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \right\rangle_{q(A, \Lambda_x)} \\
&\quad + \left\langle -\frac{pT}{2} \ln 2\pi - \frac{T}{2} \ln |\Lambda_y^{-1}| - \frac{1}{2} \sum_{t=1}^T (y_t - Bx_t)^T \Lambda_y (y_t - Bx_t) \right\rangle_{q(B, \Lambda_y)} \\
&= -\frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\Lambda_1^{-1}| - \frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \\
&\quad - \frac{k(T-1)}{2} \ln 2\pi - \left\langle \frac{T-1}{2} \ln |\Lambda_x^{-1}| \right\rangle_{q(A, \Lambda_x)} - \frac{1}{2} \sum_{t=2}^T \langle (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \rangle_{q(A, \Lambda_x)} \\
&\quad - \frac{pT}{2} \ln 2\pi - \left\langle \frac{T}{2} \ln |\Lambda_y^{-1}| \right\rangle_{q(B, \Lambda_y)} - \frac{1}{2} \sum_{t=1}^T \langle (y_t - Bx_t)^T \Lambda_y (y_t - Bx_t) \rangle_{q(B, \Lambda_y)}
\end{aligned} \tag{6.179}$$

Summarizing all terms independent of $x_1, \dots, x_T$ as a constant $C \in \mathbb{R}$ yields

$$\begin{aligned}
\langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} &= -\frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \\
&\quad - \frac{1}{2} \sum_{t=2}^T \langle (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \rangle_{q(A, \Lambda_x)}
\end{aligned}$$

$$\begin{aligned}
& -\frac{1}{2} \sum_{t=1}^T \langle (y_t - Bx_t)^T \Lambda_y (y_t - Bx_t) \rangle_{q(B, \Lambda_y)} \\
& + C
\end{aligned} \tag{6.180}$$

The right-hand side of (6.180) may be decomposed into the contribution of the exponent of an LGSSM with averaged parameters and additional “fluctuation” terms. Exemplarily, we derive this for a single term in the sum over $t = 2, \dots, T$ on the right-hand side of (6.180) and generalize the result to the remaining terms below. Specifically, we have

$$\begin{aligned}
& \langle (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \rangle_{q(A, \Lambda_x)} \\
& = \langle x_t^T \Lambda_x x_t - x^T \Lambda_x Ax_{t-1} - x_{t-1}^T A^T \Lambda_x x_t + x_{t-1}^T A^T \Lambda_x Ax_{t-1} \rangle_{q(A, \Lambda_x)} \\
& = \langle x_t^T \Lambda_x x_t \rangle_{q(A, \Lambda_x)} - \langle x_t^T \Lambda_x Ax_{t-1} \rangle_{q(A, \Lambda_x)} - \langle x_{t-1}^T A^T \Lambda_x x_t \rangle_{q(A, \Lambda_x)} + \langle x_{t-1}^T A^T \Lambda_x Ax_{t-1} \rangle_{q(A, \Lambda_x)} \\
& = x_t^T \langle \Lambda_x \rangle_{q(A, \Lambda_x)} x_t - x_t^T \langle \Lambda_x A \rangle_{q(A, \Lambda_x)} x_{t-1} - x_{t-1}^T \langle A^T \Lambda_x \rangle_{q(A, \Lambda_x)} x_t + x_{t-1}^T \langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} x_{t-1} \\
& = x_t^T \langle \Lambda_x \rangle_{q(\Lambda_x)} x_t - x_t^T \langle \Lambda_x A \rangle_{q(A|\Lambda_x)q(\Lambda_x)} x_{t-1} - x_{t-1}^T \langle A^T \Lambda_x \rangle_{q(A|\Lambda_x)q(\Lambda_x)} x_t + x_{t-1}^T \langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} x_{t-1}
\end{aligned} \tag{6.181}$$

On the right-hand side of (6.181) the first three terms involve expectations over first-order terms (or, more formally, expectations over random variables under the identity function), which may be evaluated to

$$\begin{aligned}
& \langle (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \rangle_{q(A, \Lambda_x)} \\
& = x_t^T \langle \Lambda_x \rangle_{q(\Lambda_x)} x_t - x^T \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} x_{t-1} - x_{t-1}^T \langle A \rangle_{q(A|\Lambda_x)}^T \langle \Lambda_x \rangle_{q(\Lambda_x)} x_t + x_{t-1}^T \langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} x_{t-1}
\end{aligned} \tag{6.182}$$

The last term on the right-hand side of (6.181) and (6.182) involves an expectation over a second-order term in A . For general functions g (such as the square) of a random variable, one has $\mathbb{E}(g(z)) \neq g(\mathbb{E}(z))$. Thus, to avoid the explicit evaluation of the expectation $\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)}$, we rewrite the last term of (6.182) as follows

$$\begin{aligned}
& x_{t-1}^T \langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} x_{t-1} \\
& = x_{t-1}^T \left(\langle A^T \Lambda_x A \rangle_{q(A, \Sigma_x)} + \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Sigma_x)} \langle A^T \rangle_{q(A|\Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A^T \rangle_{q(A|\Lambda_x)} \right) x_{t-1}
\end{aligned} \tag{6.183}$$

i.e., by adding and subtracting the term $x_{t-1}^T \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A^T \rangle_{q(A|\Lambda_x)} x_{t-1}$. From (6.181) and (6.182), we then have

$$\begin{aligned}
& \langle (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \rangle_{q(A, \Lambda_x)} = x_t^T \langle \Lambda_x \rangle_{q(\Lambda_x)} x_t \\
& \quad - x^T \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} x_{t-1} - x_{t-1}^T \langle A \rangle_{q(A|\Lambda_x)}^T \langle \Lambda_x \rangle_{q(\Lambda_x)} x_t \\
& \quad + x_{t-1}^T \left(\langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} \right) x_{t-1} \\
& \quad + x_{t-1}^T \left(\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} \right) x_{t-1}
\end{aligned} \tag{6.184}$$

which we may rewrite as

$$\begin{aligned} \langle (x_t - Ax_{t-1})^T \Lambda_x (x_t - Ax_{t-1}) \rangle_{q(A, \Lambda_x)} &= (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_x)} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \\ &\quad + x_{t-1}^T \left(\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} \right) x_{t-1} \end{aligned} \quad (6.185)$$

The first term in (6.185) is a quadratic term with “parameter expectations” and the second term forms the basis for the fluctuation terms. Generalizing the logic of (6.181) – (6.185) to all terms on the right-hand side of (6.180) then yields

$$\begin{aligned} \langle \ln(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} &= -\frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\Lambda_1^{-1}| - \frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \\ &\quad - \frac{1}{2} \sum_{t=2}^T (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_x)} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \\ &\quad - \frac{1}{2} \sum_{t=2}^T x_{t-1}^T \left(\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} \right) x_{t-1} \\ &\quad - \frac{1}{2} \sum_{t=1}^T (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t)^T \langle \Lambda_y \rangle_{q(\Lambda_y)} (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t) \\ &\quad - \frac{1}{2} \sum_{t=1}^T x_t^T \left(\langle B^T \Lambda_y B \rangle_{q(B, \Lambda_y)} - \langle B^T \rangle_{q(B|\Lambda_y)} \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)} \right) x_t \\ &\quad + C \end{aligned} \quad (6.186)$$

Re-ordering of the terms in (6.186) then yields

$$\begin{aligned} \langle \ln(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} &= -\frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\Lambda_1^{-1}| - \frac{1}{2} (x_1 - \mu_1)^T \Lambda_1 (x_1 - \mu_1) \\ &\quad - \frac{1}{2} \sum_{t=2}^T (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_x)} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \\ &\quad - \frac{1}{2} \sum_{t=1}^T (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t)^T \langle \Lambda_y \rangle_{q(\Lambda_y)} (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t) \\ &\quad + C \\ &\quad - \frac{1}{2} \sum_{t=2}^T x_{t-1}^T \left(\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} \right) x_{t-1} \\ &\quad - \frac{1}{2} \sum_{t=1}^T x_t^T \left(\langle B^T \Lambda_y B \rangle_{q(B, \Lambda_y)} - \langle B^T \rangle_{q(B|\Lambda_y)} \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)} \right) x_t \end{aligned} \quad (6.187)$$

From (6.180), the first four terms in (6.187) may be summarized as $\ln p(x_{1:T}, y_{1:T} | \langle \theta \rangle_{q(\theta)})$. Changing the subscript index range for the first term fluctuation term and defining

$$F_{A, \Lambda_x} := -\frac{1}{2} \sum_{t=1}^{T-1} x_t^T \left(\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} \right) x_t \quad (6.189)$$

and

$$F_{B, \Lambda_y} := -\frac{1}{2} \sum_{t=1}^T x_t^T \left(\langle B^T \Lambda_y B \rangle_{q(B, \Lambda_y)} - \langle B^T \rangle_{q(B|\Lambda_y)} \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)} \right) x_t \quad (6.190)$$

and reintroducing the superscript notation $q^{(i+1)}(\theta)$ for the variational distribution then completes the proof. $\square$

6.3.2.2 The unified inference theorem

Based on the mean and fluctuation decomposition theorem, the unified inference theorem allows for inferring $q(x_{1:T})$ in the case of random θ governed by a variational distribution $q(\theta)$. For ease of notation, we omit the superscripts $(i+1)$ for $q(x_{1:T})$ and (i) for $q(\theta)$ in the current section as above. Note that we formulate the augmented LGSSM in terms of covariance, rather than precision, matrices for consistency with Section 6.3.1.

The unified inference theorem

To infer the distribution $q(x_{1:T})$ using classical inference algorithms define an LGSSM

$$\tilde{p}_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) = p_{\mu_1, \Sigma_1}(x_1) \prod_{t=2}^T \tilde{p}_{\tilde{A}, \tilde{\Sigma}_x}(x_t | x_{t-1}) \prod_{t=1}^T \tilde{p}_{\tilde{B}, \tilde{\Sigma}_y}(\tilde{y}_t | x_t) \quad (6.191)$$

with non-random parameter $\tilde{\theta} := \{\tilde{A}, \tilde{\Sigma}_x, \tilde{B}, \tilde{\Sigma}_y\}$ for which the equality

$$\tilde{p}_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) = \langle \ln p(y_{1:T}, x_{1:T}, \theta) \rangle_{q(\theta)} \quad (6.192)$$

holds. To this end, set

$$\tilde{y}_t := \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} (t = 1, \dots, T), \tilde{A} := \langle A \rangle_{q(A|\Lambda_x)}, \tilde{\Sigma}_x := (\langle \Lambda_x \rangle_{q(\Lambda_x)})^{-1} \quad (6.193)$$

and

$$\tilde{B}_t := \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ U_A \\ U_B \end{pmatrix} (t = 1, \dots, T-1), \tilde{B}_T := \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ 0 \\ U_B \end{pmatrix} \quad (6.194)$$

with U_A and U_B defined by the Cholesky decompositions of

$$U_A^T U_A := \langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)} \quad (6.195)$$

$$U_B^T U_B := \langle B^T \Lambda_y B \rangle_{q(B, \Lambda_y)} - \langle B^T \rangle_{q(B|\Lambda_y)} \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)} \quad (6.196)$$

and

$$\tilde{\Sigma}_y := \begin{pmatrix} (\langle \Lambda_y \rangle_{q(\Lambda_y)})^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix} \quad (6.197)$$

Inferring $\tilde{p}_{\tilde{\theta}}(x_{1:T} | \tilde{y}_{1:T})$ from $\tilde{p}_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T})$ by means of KRTS smoothing then allows for the variational update

$$q(x_{1:T}) \propto \exp(\langle \ln(p(x_{1:T} | y_{1:T}, \theta)) \rangle_{q(\theta)}) = \exp(\ln \tilde{p}_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T})) = \tilde{p}_{\tilde{\theta}}(x_{1:T} | \tilde{y}_{1:T}) \quad (6.200)$$

where the proportionality statement is fulfilled with proportionality constant 1, because $\tilde{p}_{\tilde{\theta}}(x_{1:T} | \tilde{y}_{1:T})$ corresponds to a probability density function, which integrates to unity.

○ Proof of the unified inference theorem

We verify that

$$\ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) = \langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} \quad (6.201)$$

by substitution. We have

$$\begin{aligned} \ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) &= \ln \left(p_{\mu_1, \Sigma_1}(x_1) \prod_{t=2}^T p_{\tilde{A}, \tilde{\Sigma}_x}(x_t | x_{t-1}) \prod_{t=1}^T p_{\tilde{B}, \tilde{\Sigma}_y}(\tilde{y}_t | x_t) \right) \\ &= \ln \left(N(x_1; \mu_1, \Sigma_1) \prod_{t=2}^T N(x_t; \tilde{A}x_{t-1}, \tilde{\Sigma}_x) \prod_{t=1}^T N(\tilde{y}_t; \tilde{B}x_t, \tilde{\Sigma}_y) \right) \\ &= -\frac{k}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_1| - \frac{1}{2} (x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1) \\ &\quad - \frac{k(T-1)}{2} \ln 2\pi - \frac{T-1}{2} \ln |\tilde{\Sigma}_x| - \frac{1}{2} \sum_{t=2}^T \left((x_t - \tilde{A}x_{t-1})^T \tilde{\Sigma}_x^{-1} (x_t - \tilde{A}x_{t-1}) \right) \\ &\quad - \frac{(p+2k)T}{2} \ln 2\pi - \frac{T}{2} \ln |\tilde{\Sigma}_y| - \frac{1}{2} \sum_{t=1}^T \left((y_t - \tilde{B}x_t)^T \tilde{\Sigma}_y^{-1} (y_t - \tilde{B}x_t) \right) \end{aligned} \quad (6.202)$$

Summarizing all terms independent of $x_{1:T}$ in a constant $C \in \mathbb{R}$, we obtain

$$\begin{aligned} \ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) &= -\frac{1}{2} (x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1) \\ &\quad - \frac{1}{2} \sum_{t=2}^T (x_t - \tilde{A}x_{t-1})^T \tilde{\Sigma}_x^{-1} (x_t - \tilde{A}x_{t-1}) \\ &\quad - \frac{1}{2} \sum_{t=1}^T (y_t - \tilde{B}x_t)^T \tilde{\Sigma}_y^{-1} (y_t - \tilde{B}x_t) \\ &\quad + C \end{aligned} \quad (6.203)$$

Substitution of the augmented observed variable and parameters yields

$$\begin{aligned} \ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) &= -\frac{1}{2} (x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1) \\ &\quad - \frac{1}{2} \sum_{t=2}^T \left((x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_x)} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \right) \\ &\quad - \frac{1}{2} \sum_{t=1}^{T-1} \left(\begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} - \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ U_A \\ U_B \end{pmatrix} x_t \right)^T \begin{pmatrix} \langle \Lambda_y \rangle_{q(\Lambda_y)}^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} - \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ U_A \\ U_B \end{pmatrix} x_t \right) \\ &\quad - \frac{1}{2} \left(\begin{pmatrix} y_T \\ 0_k \\ 0_k \end{pmatrix} - \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ 0 \\ U_B \end{pmatrix} x_T \right)^T \begin{pmatrix} \langle \Lambda_y \rangle_{q(\Lambda_y)}^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} y_T \\ 0_k \\ 0_k \end{pmatrix} - \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ 0 \\ U_B \end{pmatrix} x_T \right) \\ &\quad + C \end{aligned} \quad (6.204)$$

Resolving the brackets of the third and fourth sum terms of the right-hand side of (6.204) then yields

$$\ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) = -\frac{1}{2} (x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1)$$

$$\begin{aligned}
& -\frac{1}{2} \sum_{t=2}^T (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_{xx})} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \\
& -\frac{1}{2} \sum_{t=1}^{T-1} \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix}^T \begin{pmatrix} (\langle \Lambda_y \rangle_{q(\Lambda_y)})^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix} \\
& + \sum_{t=1}^{T-1} \begin{pmatrix} y_t \\ 0_k \\ 0_k \end{pmatrix}^T \begin{pmatrix} (\langle \Lambda_y \rangle_{q(\Lambda_y)})^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ U_A \\ U_B \end{pmatrix} x_t \\
& -\frac{1}{2} \sum_{t=1}^{T-1} x_t^T \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ U_A \\ U_B \end{pmatrix} \begin{pmatrix} (\langle \Lambda_y \rangle_{q(\Lambda_y)})^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ U_A \\ U_B \end{pmatrix} x_t \\
& -\frac{1}{2} \begin{pmatrix} y_T \\ 0_k \\ 0_k \end{pmatrix}^T \begin{pmatrix} (\langle \Lambda_y \rangle_{q(\Lambda_y)})^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} y_T \\ 0_k \\ 0_k \end{pmatrix} \\
& + \begin{pmatrix} y_T \\ 0_k \\ 0_k \end{pmatrix}^T \begin{pmatrix} (\langle \Lambda_y \rangle_{q(\Lambda_y)})^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ 0 \\ U_B \end{pmatrix} x_T \\
& -\frac{1}{2} x_T^T \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ 0 \\ U_B \end{pmatrix} \begin{pmatrix} (\langle \Lambda_y \rangle_{q(\Lambda_y)})^{-1} & 0_{p \times k} & 0_{p \times k} \\ 0_{k \times p} & I_k & 0_{k \times k} \\ 0_{k \times p} & 0_{k \times k} & I_k \end{pmatrix}^{-1} \begin{pmatrix} \langle B \rangle_{q(B|\Lambda_y)} \\ 0 \\ U_B \end{pmatrix} x_T \\
& + C
\end{aligned} \tag{6.205}$$

Writing out the matrix products on the right-hand side of (6.205) then yields

$$\begin{aligned}
\ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) &= -\frac{1}{2} (x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1) \\
& -\frac{1}{2} \sum_{t=2}^T (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_{xx})} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \\
& -\frac{1}{2} \sum_{t=1}^{T-1} y_t \langle \Lambda_y \rangle_{q(\Lambda_y)} y_t \\
& + \sum_{t=1}^{T-1} y_t \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)} x_t \\
& -\frac{1}{2} \sum_{t=1}^{T-1} x_t^T \langle B \rangle_{q(B|\Lambda_y)}^T \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)} x_t \\
& -\frac{1}{2} \sum_{t=1}^{T-1} x_t^T U_A^T U_A x_t \\
& -\frac{1}{2} \sum_{t=1}^{T-1} x_t^T U_B^T U_B x_t \\
& -\frac{1}{2} y_T \langle \Lambda_y \rangle_{q(\Sigma_y)} y_T
\end{aligned}$$

$$\begin{aligned}
& + y_T \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)} x_T \\
& - \frac{1}{2} x_T^T U_B^T U_B x_T \\
& + C
\end{aligned} \tag{6.206}$$

Summarizing terms on the right-hand side of (6.206) and writing out the matrix products $U_A^T U_A$ and $U_B^T U_B$ then gives

$$\begin{aligned}
\ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) = & -\frac{1}{2} (x_1 - \mu_1)^T \Sigma_1^{-1} (x_1 - \mu_1) \\
& - \frac{1}{2} \sum_{t=2}^T (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1})^T \langle \Lambda_x \rangle_{q(\Lambda_x)} (x_t - \langle A \rangle_{q(A|\Lambda_x)} x_{t-1}) \\
& - \frac{1}{2} \sum_{t=1}^T (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t)^T \langle \Lambda_y \rangle_{q(\Lambda_y)} (y_t - \langle B \rangle_{q(B|\Lambda_y)} x_t) \\
& - \frac{1}{2} \sum_{t=1}^{T-1} x_t^T (\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)}) x_t \\
& - \frac{1}{2} \sum_{t=1}^T x_t^T (\langle B^T \Lambda_y B \rangle_{q(B, \Lambda_y)} - \langle B^T \rangle_{q(B|\Lambda_y)} \langle \Lambda_y \rangle_{q(\Lambda_y)} \langle B \rangle_{q(B|\Lambda_y)}) x_t \\
& + C
\end{aligned} \tag{6.207}$$

The first three terms on the right-hand side of (6.207) together with the appropriate terms in the constant C may now be recognized as an LGSSM with “averaged parameters” and we have

$$\begin{aligned}
\ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) = & \ln p(x_{1:T}, y_{1:T} | \langle \theta \rangle_{q(\theta)}) \\
& - \frac{1}{2} \sum_{t=1}^{T-1} x_t^T (\langle A^T \Lambda_x A \rangle_{q(A, \Lambda_x)} - \langle A^T \rangle_{q(A|\Lambda_x)} \langle \Lambda_x \rangle_{q(\Lambda_x)} \langle A \rangle_{q(A|\Lambda_x)}) x_t \\
& - \frac{1}{2} \sum_{t=1}^T x_t^T (\langle B^T \Sigma_y^{-1} B \rangle_{q(B, \Sigma_y)} - \langle B^T \rangle_{q(B|\Sigma_y)} \langle \Sigma_y^{-1} \rangle_{q(\Sigma_y)} \langle B \rangle_{q(B|\Sigma_y)}) x_t
\end{aligned} \tag{6.208}$$

For $\Lambda_1 = \Sigma_1^{-1}$ The right-hand side of (6.208) is identical to (6.171) and we thus have shown that

$$\ln p_{\tilde{\theta}}(\tilde{y}_{1:T}, x_{1:T}) = \langle \ln p(y_{1:T}, x_{1:T} | \theta) \rangle_{q(\theta)} \tag{6.209}$$

if $\tilde{\theta}$ and $\tilde{y}_{1:T}$ are defined as in (6.193) – (6.196). This concludes the proof. $\square$

6.3.2.3 Unified inference for the tutorial example

We now consider the variational distribution update equation (6.21) for the tutorial example

$$q^{(i+1)}(x_{1:T}) \propto \exp \left(\langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i+1)}(a) q^{(i+1)}(\lambda_x) q^{(i+1)}(b) q^{(i+1)}(\lambda_y)} \right) \tag{6.210}$$

where we set according to the unified inference theorem

$$q^{(i+1)}(x_{1:T}) \propto \exp \left(\langle \ln p(x_{1:T} | y_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q(a, \lambda_x, b, \lambda_y)} \right) = \exp(\ln \tilde{p}_{\tilde{\theta}^{(i)}}(\tilde{y}_{1:T}, x_{1:T})) = \tilde{p}_{\tilde{\theta}^{(i)}}(x_{1:T} | \tilde{y}_{1:T}) \tag{6.211}$$

Applying the observed variable and parameter augmentations introduced in the unified inference theorem to the univariate LGSSM under consideration in the tutorial, we obtain

$$\tilde{y}_t := \begin{pmatrix} y_t \\ 0 \\ 0 \end{pmatrix} \text{ for } t = 1, \dots, T \quad (6.214)$$

$$\tilde{a}^{(i)} := \langle a \rangle_{q^{(i+1)}(a)} = \mu_a^{(i+1)} \quad (6.215)$$

$$\tilde{\sigma}_x^{2(i)} := \left(\langle \lambda_x \rangle_{q^{(i+1)}(\lambda_x)} \right)^{-1} = \left(a_{\lambda_x}^{(i+1)} b_{\lambda_x}^{(i+1)} \right)^{-1} \quad (6.216)$$

$$\begin{aligned} \tilde{b}_t^{(i)} &:= \begin{pmatrix} \langle b \rangle_{q^{(i+1)}(b)} \\ u_A \\ u_B \end{pmatrix} \\ &= \begin{pmatrix} \langle b \rangle_{q(b)} \\ \sqrt{\frac{\langle a^2 \lambda_x \rangle_{q^{(i+1)}(a)q^{(i+1)}(\lambda_x)} - \langle a \rangle_{q^{(i+1)}(a)}^2 \langle \lambda_x \rangle_{q^{(i+1)}(\lambda_x)}}{\langle b^2 \lambda_y \rangle_{q^{(i+1)}(b)q^{(i+1)}(\lambda_y)} - \langle b \rangle_{q^{(i+1)}(b)}^2 \langle \lambda_y \rangle_{q^{(i+1)}(\lambda_y)}}} \\ \mu_b^{(i+1)} \\ \sqrt{\frac{(\mu_a^{(i+1)^2} + \sigma_a^{2(i+1)}) a_{\lambda_x}^{(i+1)} b_{\lambda_x}^{(i+1)} - \mu_a^{(i+1)^2} a_{\lambda_x}^{(i+1)} b_{\lambda_x}^{(i+1)}}{(\mu_b^{(i+1)^2} + \sigma_b^{2(i+1)}) a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)} - \mu_b^{(i+1)^2} a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)}}} \end{pmatrix} \\ &= \begin{pmatrix} \mu_b^{(i+1)} \\ \sqrt{\frac{\sigma_a^{2(i+1)} a_{\lambda_x}^{(i+1)} b_{\lambda_x}^{(i+1)}}{\sigma_b^{2(i+1)} a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)}}} \\ \sqrt{\frac{\sigma_a^{2(i+1)} a_{\lambda_x}^{(i+1)} b_{\lambda_x}^{(i+1)}}{\sigma_b^{2(i+1)} a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)}}} \end{pmatrix} \quad \text{for } t = 1, \dots, T-1 \end{aligned} \quad (6.218)$$

$$\tilde{b}_T^{(i)} := \begin{pmatrix} \mu_b^{(i+1)} \\ 0 \\ \sqrt{\frac{\sigma_b^{2(i+1)} a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)}}{\sigma_b^{2(i+1)} a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)}}} \end{pmatrix} \quad (6.219)$$

and

$$\tilde{\sigma}_y^{2(i)} := \begin{pmatrix} \left(\langle \lambda_y \rangle_{q^{(i+1)}(\lambda_y)} \right)^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \left(a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)} \right)^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (6.220)$$

The KRTS inference machinery then allows for inferring the expectation and covariance parameters of

$$q^{(i)}(x_{1:T}) = \tilde{p}_{\tilde{\theta}^{(i)}}(x_{1:T} | \tilde{y}_{1:T}) = \tilde{p}_{\tilde{\theta}^{(i)}}(x_1 | \tilde{y}_{1:T}) \prod_{t=2}^T \frac{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}, x_t | \tilde{y}_{1:T})}{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1} | \tilde{y}_{1:T})} \quad (6.221)$$

The distributions on the right-hand side of (6.221) are given in parameterized form by

$$\tilde{p}_{\tilde{\theta}^{(i)}}(x_t|\tilde{y}_{1:T}) = N\left(x_t; \tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)}, \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)}\right) \quad (6.222)$$

for $t = 1, \dots, T$, and

$$\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}, x_t|\tilde{y}_{1:T}) = N\left(x_t; \tilde{\mu}_{x_{t-1:t}|\tilde{y}_{1:T}}^{(i)}, \tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^{(i)}\right) \quad (6.223)$$

for $t = 2, \dots, T$, respectively. Note that we use the tilde notation throughout, to distinguish the parameters of these distributions from the parameters inferred from a non-augmented LGSSM by means of KRTS smoothing. Consequently the remaining expectations in the variational update equations for a, λ_x, b and λ_y shown in Table 5 evaluate in parameterized form to

$$\langle x_t^2 \rangle_{q^{(i+1)}(x_t)} = \langle x_t^2 \rangle_{p_{\tilde{\theta}^{(i)}}(x_t|\tilde{y}_{1:T})} = \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)}\right)^2 \quad (6.224)$$

for $t = 1, \dots, T$, and

$$\langle x_{t-1}x_t \rangle_{q^{(i+1)}(x_{t-1:t})} = \langle x_t x_{t-1} \rangle_{p_{\tilde{\theta}^{(i)}}(x_{t-1:t}|\tilde{y}_{1:T})} = \tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^2{}^{(i)} + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)} \quad (6.222)$$

for $t = 2, \dots, T$, where $\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^2{}^{(i)}$ denotes the off-diagonal element in the first row of $\tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^{(i)}$. Table 6 summarizes the development of the variational update and unified inference framework thus far.

% initialization of the variational distribution parameters for $a, \sigma_x^2, b, \sigma_y^2, x_{1:T}$

$$\sigma_a^{2(0)} := \sigma_a^2, \mu_a^{(0)} := \mu_a, a_{\lambda_x}^{(0)} := a_{\lambda_x}, b_{\lambda_x}^{(0)} := b_{\lambda_x}, \sigma_b^{2(0)} := \sigma_b^2, \mu_b^{(0)} := \mu_b, a_{\lambda_y}^{(0)} := a_{\lambda_y}, b_{\lambda_y}^{(0)} := b_{\lambda_y}$$

$$\tilde{y}_t := \begin{pmatrix} y_t \\ 0 \\ 0 \end{pmatrix}, \tilde{a}^{(0)} := \mu_a^{(0)}, \tilde{\sigma}_x^{2(0)} := (a_{\lambda_x}^{(0)} b_{\lambda_x}^{(0)})^{-1}, \tilde{b}_t^{(0)} := \begin{pmatrix} \mu_b^{(0)} \\ \sqrt{\sigma_a^{2(0)} a_{\lambda_x}^{(0)} b_{\lambda_x}^{(0)}} \\ \sqrt{\sigma_b^{2(0)} a_{\lambda_y}^{(0)} b_{\lambda_y}^{(0)}} \end{pmatrix}, \tilde{b}_T^{(0)} := \begin{pmatrix} \mu_b^{(0)} \\ 0 \\ \sqrt{\sigma_b^{2(0)} a_{\lambda_y}^{(0)} b_{\lambda_y}^{(0)}} \end{pmatrix}, \tilde{\sigma}_y^{2(0)} := \begin{pmatrix} (a_{\lambda_y}^{(0)} b_{\lambda_y}^{(0)})^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\left(\tilde{\mu}_{x_t|\tilde{y}_{1:t}}^{(0)}, \tilde{\sigma}_{x_t|\tilde{y}_{1:t}}^{2(0)} \right)_{t=1,\dots,T} := \text{Kalman Filter} \left(y_{1:T}, \mu_1, \sigma_1^2, \tilde{a}^{(0)}, \tilde{\sigma}_x^{2(0)}, \tilde{b}_t^{(0)}, \tilde{\sigma}_y^{2(0)} \right)$$

$$\left(\left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(0)}, \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^{2(0)} \right)_{t=1,\dots,T}, \left(\tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^{(0)} \right)_{t=2,\dots,T} \right) := \text{KRTS Smoother} \left(y_{1:T}, \tilde{\mu}_{x_t|\tilde{y}_{1:t}}^{(0)}, \tilde{a}^{(0)}, \tilde{\sigma}_{x_t|\tilde{y}_{1:t}}^{2(0)}, \tilde{\sigma}_x^{2(0)} \right)$$

% iterative update equations for the variational distribution parameters

for $i = 0, 1, 2 \dots$

$$\sigma_a^{2(i+1)} := \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^{2(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) + \frac{1}{\sigma_a^2} \right)^{-1}$$

$$\mu_a^{(i+1)} := \sigma_a^{2(i+1)} \left(a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \left(\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^{2(i)} + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)} \right) + \frac{\mu_a}{\sigma_a^2} \right)$$

$$a_{\lambda_x}^{(i+1)} := \frac{T}{2} + a_{\lambda_x}$$

$$b_{\lambda_x}^{(i+1)} := \left(\frac{1}{b_{\lambda_x}} + \frac{1}{2} \left(\sum_{t=2}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^{2(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) - 2 \sum_{t=2}^T \mu_a^{(i+1)} \left(\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^{2(i)} + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)} \right) + \left(\sigma_a^{2(i+1)} + \mu_a^{(i+1)^2} \right) \sum_{t=2}^T \left(\tilde{\sigma}_{x_{t-1}|\tilde{y}_{1:T}}^{2(i)} + \left(\tilde{\mu}_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) \right) \right)^{-1}$$

$$\sigma_b^{2(i+1)} := \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^{2(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) + \frac{1}{\sigma_b^2} \right)^{-1}$$

$$\mu_b^{(i+1)} := \sigma_b^{2(i+1)} \left(a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \sum_{t=1}^T y_t \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^{2(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) + \frac{\mu_b}{\sigma_b^2} \right)$$

$$a_{\lambda_y}^{(i+1)} := \frac{T}{2} + a_{\lambda_y}$$

$$b_{\lambda_y}^{(i+1)} := \left(\frac{1}{b_{\lambda_y}} + \frac{1}{2} \left(\sum_{t=1}^T y_t^2 - 2 \mu_b^{(i+1)} \sum_{t=1}^T y_t \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^{2(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) + \left(\sigma_b^{2(i+1)} + \mu_b^{(i+1)^2} \right) \sum_{t=1}^T \left(\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^{2(i)} + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)} \right) \right) \right)^{-1}$$

$$\tilde{y}_t := \begin{pmatrix} y_t \\ 0 \\ 0 \end{pmatrix}, \tilde{a}^{(i)} := \mu_a^{(i+1)}, \tilde{\sigma}_x^{2(i)} := (a_{\lambda_x}^{(i+1)} b_{\lambda_x}^{(i+1)})^{-1}, \tilde{b}_t^{(i)} := \begin{pmatrix} \mu_b^{(i+1)} \\ \sqrt{\sigma_a^{2(i+1)} a_{\lambda_x}^{(i+1)} b_{\lambda_x}^{(i+1)}} \\ \sqrt{\sigma_b^{2(i+1)} a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)}} \end{pmatrix}, \tilde{b}_T^{(i)} := \begin{pmatrix} \mu_b^{(i+1)} \\ 0 \\ \sqrt{\sigma_b^{2(i+1)} a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)}} \end{pmatrix}, \tilde{\sigma}_y^{2(i)} := \begin{pmatrix} (a_{\lambda_y}^{(i+1)} b_{\lambda_y}^{(i+1)})^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\left(\tilde{\mu}_{x_t|\tilde{y}_{1:t}}^{(i+1)}, \tilde{\sigma}_{x_t|\tilde{y}_{1:t}}^{2(i+1)} \right)_{t=1,\dots,T} := \text{Kalman Filter} \left(y_{1:T}, \mu_1, \sigma_1^2, \tilde{a}^{(i)}, \tilde{\sigma}_x^{2(i)}, \tilde{b}_t^{(i)}, \tilde{\sigma}_y^{2(i)} \right)$$

$$\left(\left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i+1)}, \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^{2(i+1)} \right)_{t=1,\dots,T}, \left(\tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^{(i+1)} \right)_{t=2,\dots,T} \right) := \text{KRTS Smoother} \left(y_{1:T}, \tilde{\mu}_{x_t|\tilde{y}_{1:t}}^{(i)}, \tilde{a}^{(i)}, \tilde{\sigma}_{x_t|\tilde{y}_{1:t}}^{2(i)}, \tilde{\sigma}_x^{2(i)} \right)$$

Table 6. Pseudocode for parameterized variational parameter update equations for the unobserved variables $a, b, \lambda_x, \lambda_y, x_{1:T}$

6.4 Evaluation of the variational free energy $\mathcal{F}(q(\vartheta))$

In Supplement Section 5.5, we rewrote the variational free energy as the sum of an “average likelihood” and the KL divergence between the variational approximation to the posterior distribution over unobserved variables and the marginal (prior) over unobserved variables as follows:

$$\mathcal{F}(q^{(i)}(\vartheta)) = \mathcal{L}_{av}(p(y|\vartheta), q^{(i)}(\vartheta)) - \mathcal{KL}(q^{(i)}(\vartheta) || p(\vartheta)) = \int q^{(i)}(\vartheta) \ln p(y|\vartheta) d\vartheta - \int q^{(i)}(\vartheta) \ln \left(\frac{q^{(i)}(\vartheta)}{p(\vartheta)} \right) d\vartheta \quad (6.223)$$

Using the variational free energy decomposition of (6.223) here would necessitate (1) the explicit derivation of the “average likelihood term” for the current model under consideration, and (2) the explicit derivation of the marginal distribution $p(x_{1:T})$ of in the joint distribution specified by (6.1) based on its conditional independence properties and the functional forms of its subcomponents specified by (6.2) – (6.10). By using an alternative decomposition of the variational free energy, we can circumvent these derivations, and capitalize on variational expectation of $p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y)$ derived in Section 6.2 and the known differential entropies of the distributions comprising the variational approximation. Specifically, we here use the following decomposition of the variational free energy:

$$\mathcal{F}(q^{(i)}(\vartheta)) = \mathcal{E}_{av}(p(y, \vartheta), q^{(i)}(\vartheta)) + \mathcal{H}(q^{(i)}(\vartheta)) = \int q^{(i)}(\vartheta) \ln(p(y, \vartheta)) d\vartheta - \int q^{(i)}(\vartheta) \ln(q^{(i)}(\vartheta)) d\vartheta \quad (6.224)$$

where we refer to the terms $\mathcal{E}_{av}(p(y, \vartheta), q^{(i)}(\vartheta))$ and $\mathcal{H}(q^{(i)}(\vartheta))$ as “average energy” and entropy, respectively. For clarity, we derive both decompositions below, omitting the algorithm iteration superscript (i) of the variational distribution for ease of notation.

○ Derivation of (6.223) and (6.224)

By definition (cf. equation (13) of the tutorial), we have

$$\mathcal{F}(q(\vartheta)) := \int q(\vartheta) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta)} \right) d\vartheta \quad (6.225)$$

Rewriting the joint distribution over observed and unobserved variables as

$$p(y, \vartheta) = p(y|\vartheta)p(\vartheta) \quad (6.225)$$

results in the “average likelihood – KL divergence” decomposition of the variational free energy as follows:

$$\begin{aligned} \mathcal{F}(q(\vartheta)) &:= \int q(\vartheta) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta)} \right) d\vartheta \\ &= \int q(\vartheta) \ln \left(\frac{p(y|\vartheta)p(\vartheta)}{q(\vartheta)} \right) d\vartheta \\ &= \int q(\vartheta) \left(\ln p(y|\vartheta) + \ln \left(\frac{p(\vartheta)}{q(\vartheta)} \right) \right) d\vartheta \\ &= \int q(\vartheta) \ln p(y|\vartheta) d\vartheta + \int q(\vartheta) \ln \left(\frac{p(\vartheta)}{q(\vartheta)} \right) d\vartheta \\ &= \int q(\vartheta) \ln p(y|\vartheta) d\vartheta - \int q(\vartheta) \ln \left(\frac{q(\vartheta)}{p(\vartheta)} \right) d\vartheta \\ &:= \mathcal{L}_{av}(p(y|\vartheta), q(\vartheta)) - \mathcal{KL}(q(\vartheta) || p(\vartheta)) \end{aligned} \quad (6.226)$$

Note that the KL-divergence is defined by the fact that the probability distribution integrated over corresponds to the numerator of the ratio in the argument of the logarithm (Cover and Thomas (2006), p.251). Merely decomposing

the argument of the logarithm in the definition of the variational free energy results in its “average energy - entropy” decomposition as follows

$$\begin{aligned}
\mathcal{F}(q(\vartheta)) &:= \int q(\vartheta) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta)} \right) d\vartheta \\
&= \int q(\vartheta) \ln \left(\frac{p(y, \vartheta)}{q(\vartheta)} \right) d\vartheta \\
&= \int q(\vartheta) \left(\ln p(y, \vartheta) + \ln \left(\frac{1}{q(\vartheta)} \right) \right) d\vartheta \\
&= \int q(\vartheta) \ln p(y, \vartheta) d\vartheta + \int q(\vartheta) \ln \left(\frac{1}{q(\vartheta)} \right) d\vartheta \\
&= \int q(\vartheta) \ln p(y, \vartheta) d\vartheta - \int q(\vartheta) \ln q(\vartheta) d\vartheta \\
&=: \mathcal{E}_{av} \left(p(y, \vartheta), q^{(i)}(\vartheta) \right) + \mathcal{H} \left(q^{(i)}(\vartheta) \right)
\end{aligned} \tag{6.227}$$

Note that the differential entropy is defined as

$$\mathcal{H}(q(\vartheta)) := - \int q(\vartheta) \ln q(\vartheta) d\vartheta \tag{6.228}$$

(Cover and Thomas (2006), p.243), thus the change in sign in the last definition of (6.227).

□

Based on the “average energy - entropy” decomposition, the variational free energy for the current generative model and variational approximation can be evaluated to

$$\mathcal{F} \left(q^{(i)}(\vartheta) \right) = \mathcal{E}_{av} \left(p(y, \vartheta), q^{(i)}(\vartheta) \right) + \mathcal{H} \left(q^{(i)}(\vartheta) \right) \tag{6.229}$$

where

$$\begin{aligned}
&\mathcal{E}_{av} \left(p(y, \vartheta), q^{(i)}(\vartheta) \right) \\
&= \frac{1}{2} \ln \left(\frac{\lambda_{x_1}}{2\pi} \right) - \frac{1}{2} \lambda_{x_1} \left((\tilde{\sigma}_{x_1|\tilde{y}_{1:T}}^2)^{(i)} + (\tilde{\mu}_{x_1|\tilde{y}_{1:T}}^{(i)})^2 \right) - 2\tilde{\mu}_{x_1|\tilde{y}_{1:T}}^{(i)} \mu_{x_1} + \mu_{x_1}^2 \\
&\quad + \frac{T}{2} \left(\psi(a_{\lambda_x}^{(i)}) + \ln(b_{\lambda_x}^{(i)}) \right) - \frac{T}{2} \ln(2\pi) \\
&\quad - \frac{1}{2} a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \left(\sum_{t=2}^T (\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2)^{(i)} + (\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)})^2 \right) - 2\mu_a^{(i)} \sum_{t=2}^T (\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^2 + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)}) + \left(\sigma_a^2 + (\mu_a^{(i)})^2 \right) \sum_{t=1}^{T-1} (\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2)^{(i)} + (\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)})^2 \\
&\quad + \frac{T}{2} \left(\psi(a_{\lambda_y}^{(i)}) + \ln(b_{\lambda_y}^{(i)}) \right) - \frac{T}{2} \ln(2\pi) \\
&\quad - \frac{1}{2} a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \left(\sum_{t=2}^T y_t^2 - 2\mu_b^{(i)} \sum_{t=2}^T y_t \tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} + \left(\sigma_b^2 + (\mu_b^{(i)})^2 \right) \sum_{t=1}^T (\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2)^{(i)} + (\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)})^2 \right) \\
&\quad - \frac{1}{2\sigma_a^2} \left(\left(\sigma_a^2 + (\mu_a^{(i)})^2 \right) - 2\mu_a^{(i)} \mu_a + \mu_a^2 \right) - \frac{1}{2} \ln(2\pi\sigma_a^2) \\
&\quad - \ln(\Gamma(a_{\lambda_x})) - a_{\lambda_x} \ln(b_{\lambda_x}) + (a_{\lambda_x} - 1) \left(\psi(a_{\lambda_x}^{(i)}) + \ln(b_{\lambda_x}^{(i)}) \right) - \frac{1}{b_{\lambda_x}} a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \\
&\quad - \frac{1}{2\sigma_b^2} \left(\left(\sigma_b^2 + (\mu_b^{(i)})^2 \right) - 2\mu_b^{(i)} \mu_b + \mu_b^2 \right) - \frac{1}{2} \ln 2\pi\sigma_b^2 \\
&\quad - \ln(\Gamma(a_{\lambda_y})) - a_{\lambda_y} \ln(b_{\lambda_y}) + (a_{\lambda_y} - 1) \left(\psi(a_{\lambda_y}^{(i)}) + \ln(b_{\lambda_y}^{(i)}) \right) - \frac{1}{b_{\lambda_y}} a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{H}\left(q^{(i)}(\vartheta)\right) &= \frac{1}{2} \ln\left(2\pi e \sigma_a^2{}^{(i)}\right) \\
&+ a_{\lambda_x}^{(i)} + \ln\left(b_{\lambda_x}^{(i)}\right) + \ln \Gamma\left(a_{\lambda_x}^{(i)}\right) + \left(1 - a_{\lambda_x}^{(i)}\right) \psi(a_{\lambda_x}^{(i)}) \\
&+ \frac{1}{2} \ln\left(2\pi e \sigma_b^2{}^{(i)}\right) \\
&+ a_{\lambda_y}^{(i)} + \ln\left(b_{\lambda_y}^{(i)}\right) + \ln \Gamma\left(a_{\lambda_y}^{(i)}\right) + \left(1 - a_{\lambda_y}^{(i)}\right) \psi(a_{\lambda_y}^{(i)}) \\
&+ (T-1) \ln(2\pi e) + \sum_{t=2}^T \ln |\tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}| - \frac{T-2}{2} \sum_{t=2}^{T-1} \ln \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2
\end{aligned} \tag{6.231}$$

◦ Derivation of (6.229) – (6.331)

Based on (6.224)

$$\mathcal{F}\left(q^{(i)}(\vartheta)\right) = \int q^{(i)}(\vartheta) \ln(p(y, \vartheta)) d\vartheta - \int q^{(i)}(\vartheta) \ln\left(q^{(i)}(\vartheta)\right) d\vartheta \tag{6.230}$$

we first recall the definitions of $y, \vartheta, p(y, \vartheta)$ and $q^{(i)}(\vartheta)$ for the generative model and variational approximation under consideration. As noted in (6.14), we have

$$y := y_{1:T} \text{ and } \vartheta := \{(x_{1:T}, a, \lambda_x, b, \lambda_y)\} \tag{6.231}$$

and thus

$$p(y, \vartheta) := p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \tag{6.232}$$

From (6.12), we further have

$$q^{(i)}(\vartheta) = q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T}) \tag{6.233}$$

The “average energy term” in (6.230)

$$\begin{aligned}
\int q^{(i)}(\vartheta) \ln(p(y, \vartheta)) d\vartheta &= \int q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T}) \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) da d\lambda_x db d\lambda_y dx_{1:T} \\
&= \langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T})}
\end{aligned} \tag{6.234}$$

has been evaluated in Section 6.2 for $i = j = k = l = m = n$ as

$$\begin{aligned}
&\langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T})} \\
&= \frac{1}{2} \ln\left(\frac{\lambda_{x_1}}{2\pi}\right) - \frac{1}{2} \lambda_{x_1} \left(\langle x_1^2 \rangle_{q^{(i)}(x_1)} - 2 \langle x_1 \rangle_{q^{(i)}(x_1)} \mu_{x_1} + \mu_{x_1}^2 \right) \\
&+ \frac{T}{2} \langle \ln \lambda_x \rangle_{q^{(i)}(\lambda_x)} - \frac{T}{2} \ln(2\pi) - \frac{1}{2} \langle \lambda_x (\sum_{t=2}^T x_t^2 - 2a \sum_{t=2}^T x_{t-1} x_t + a^2 \sum_{t=2}^T x_{t-1}^2) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})}
\end{aligned}$$

$$\begin{aligned}
& + \frac{T}{2} \langle \ln \lambda_y \rangle_{q^{(i)}(\lambda_y)} - \frac{T}{2} \ln(2\pi) - \frac{1}{2} \langle \lambda_y (\sum_{t=1}^T y_t^2 - 2b \sum_{t=1}^T y_t x_t + b^2 \sum_{t=1}^T x_t^2) \rangle_{q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_t)} \\
& - \frac{1}{2\sigma_a^2} \langle a^2 - 2a\mu_a + \mu_a^2 \rangle_{q^{(i)}(a)} - \frac{1}{2} \ln(2\pi\sigma_a^2) \\
& - \ln(\Gamma(a_{\lambda_x})) - a_{\lambda_x} \ln(b_{\lambda_x}) + (a_{\lambda_x} - 1) \langle \ln \lambda_x \rangle_{q^{(i)}(\lambda_x)} - \frac{1}{b_{\lambda_x}} \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \\
& - \frac{1}{2\sigma_b^2} \langle b^2 - 2b\mu_b + \mu_b^2 \rangle_{q^{(i)}(b)} - \frac{1}{2} \ln(2\pi\sigma_b^2) \\
& - \ln(\Gamma(a_{\lambda_y})) - a_{\lambda_y} \ln(b_{\lambda_y}) + (a_{\lambda_y} - 1) \langle \ln \lambda_y \rangle_{q^{(i)}(\lambda_y)} - \frac{1}{b_{\lambda_y}} \langle \lambda_y \rangle_{q^{(i)}(\lambda_y)}
\end{aligned} \tag{6.235}$$

In the following, we derive the remaining expectations to obtain an expression of the average energy in terms of the variational and prior parameters. Specifically, from Supplement 6.3.2.3 we have

$$\langle x_1^2 \rangle_{q^{(i)}(x_1)} = \tilde{\sigma}_{x_1|\tilde{y}_{1:T}}^{2(i)} + \left(\tilde{\mu}_{x_1|\tilde{y}_{1:T}}^{(i)} \right)^2 \quad \text{and} \quad \langle x_1 \rangle_{q^{(i)}(x_1)} = \tilde{\mu}_{x_1|\tilde{y}_{1:T}}^{(i)} \tag{6.236}$$

From Supplement 2.7, we have

$$\langle \ln(\lambda_x) \rangle_{q^{(i)}(\lambda_x)} = \left(\psi(a_{\lambda_x}^{(i)}) + \ln b_{\lambda_x}^{(i)} \right) \quad \text{and} \quad \langle \ln(\lambda_y) \rangle_{q^{(i)}(\lambda_y)} = \left(\psi(a_{\lambda_y}^{(i)}) + \ln b_{\lambda_y}^{(i)} \right) \tag{6.237}$$

We further have

$$\langle a^2 - 2a\mu_a + \mu_a^2 \rangle_{q^{(i)}(a)} = \langle a^2 \rangle_{q^{(i)}(a)} - 2\langle a \rangle_{q^{(i)}(a)}\mu_a + \mu_a^2 = \left(\sigma_a^{2(i)} + \left(\mu_a^{(i)} \right)^2 \right) - 2\mu_a^{(i)}\mu_a + \mu_a^2 \tag{6.238}$$

and likewise

$$\langle b^2 - 2b\mu_b + \mu_b^2 \rangle_{q^{(i)}(b)} = \langle b^2 \rangle_{q^{(i)}(b)} - 2\langle b \rangle_{q^{(i)}(b)}\mu_b + \mu_b^2 = \left(\sigma_b^{2(i)} + \left(\mu_b^{(i)} \right)^2 \right) - 2\mu_b^{(i)}\mu_b + \mu_b^2 \tag{6.239}$$

Finally, we have

$$\begin{aligned}
& \langle \lambda_x (\sum_{t=2}^T x_t^2 - 2a \sum_{t=2}^T x_{t-1}x_t + a^2 \sum_{t=2}^T x_{t-1}^2) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} \\
& = \langle (\lambda_x \sum_{t=2}^T x_t^2) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} \\
& \quad - \langle (2a\lambda_x \sum_{t=2}^T x_{t-1}x_t) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} \\
& \quad + \langle (a^2 \lambda_x \sum_{t=2}^T x_{t-1}^2) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(x_{t-1:t})} \\
& = \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \sum_{t=2}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)} \\
& \quad - 2\langle a \rangle_{q^{(i)}(a)} \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \sum_{t=2}^T \langle x_{t-1}x_t \rangle_{q^{(i)}(x_{t-1:t})} \\
& \quad + \langle a^2 \rangle_{q^{(i)}(a)} \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \sum_{t=2}^T \langle x_{t-1}^2 \rangle_{q^{(i)}(x_{t-1})} \\
& = \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \sum_{t=2}^T \langle x_t^2 \rangle_{q^{(i)}(x_t)}
\end{aligned}$$

$$\begin{aligned}
& -2\langle a \rangle_{q^{(i)}(a)} \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \sum_{t=2}^T \langle x_{t-1} x_t \rangle_{q(x_{t-1:t})} \\
& + \langle a^2 \rangle_{q^{(i)}(a)} \langle \lambda_x \rangle_{q^{(i)}(\lambda_x)} \sum_{t=2}^T \langle x_{t-1}^2 \rangle_{q(x_{t-1})} \\
& = a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) \\
& - 2\mu_a^{(i)} a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=2}^T \left(\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^2 + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)} \right) \\
& + \left(\sigma_a^2{}^{(i)} + \left(\mu_a^{(i)} \right)^2 \right) a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \sum_{t=1}^{T-1} \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) \\
& = a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \left(\sum_{t=2}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) - 2\mu_a^{(i)} \sum_{t=2}^T \left(\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^2 + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)} \right) + \left(\sigma_a^2{}^{(i)} + \left(\mu_a^{(i)} \right)^2 \right) \sum_{t=1}^{T-1} \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) \right)
\end{aligned} \tag{6.240}$$

and, following the same line of reasoning as in (6.240)

$$\begin{aligned}
& \langle \lambda_y (\sum_{t=1}^T y_t^2 - 2b \sum_{t=1}^T y_t x_t + b^2 \sum_{t=1}^T x_t^2) \rangle_{q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_t)} \\
& = a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \left(\sum_{t=1}^T y_t^2 - 2\mu_b^{(i)} \sum_{t=1}^T y_t \tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} + \left(\sigma_b^2{}^{(i)} + \left(\mu_b^{(i)} \right)^2 \right) \sum_{t=1}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) \right)
\end{aligned} \tag{6.241}$$

Substitution of (6-236) – (6.245) into (6.235) then yields

$$\begin{aligned}
& \langle \ln p(y_{1:T}, x_{1:T}, a, \lambda_x, b, \lambda_y) \rangle_{q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T})} \\
& = \frac{1}{2} \ln \left(\frac{\lambda_{x_1}}{2\pi} \right) - \frac{1}{2} \lambda_{x_1} \left(\left(\tilde{\sigma}_{x_1|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_1|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) - 2\tilde{\mu}_{x_1|\tilde{y}_{1:T}}^{(i)} \mu_{x_1} + \mu_{x_1}^2 \right) \\
& + \frac{T}{2} \left(\psi(a_{\lambda_x}^{(i)}) + \ln(b_{\lambda_x}^{(i)}) \right) - \frac{T}{2} \ln(2\pi) \\
& - \frac{1}{2} a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \left(\sum_{t=2}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) - 2\mu_a^{(i)} \sum_{t=2}^T \left(\tilde{\sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}^2 + \mu_{x_{t-1}|\tilde{y}_{1:T}}^{(i)} \mu_{x_t|\tilde{y}_{1:T}}^{(i)} \right) + \left(\sigma_a^2{}^{(i)} + \left(\mu_a^{(i)} \right)^2 \right) \sum_{t=1}^{T-1} \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) \right) \\
& + \frac{T}{2} \left(\psi(a_{\lambda_y}^{(i)}) + \ln(b_{\lambda_y}^{(i)}) \right) - \frac{T}{2} \ln(2\pi) \\
& - \frac{1}{2} a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)} \left(\sum_{t=1}^T y_t^2 - 2\mu_b^{(i)} \sum_{t=1}^T y_t \tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} + \left(\sigma_b^2{}^{(i)} + \left(\mu_b^{(i)} \right)^2 \right) \sum_{t=1}^T \left(\tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2{}^{(i)} + \left(\tilde{\mu}_{x_t|\tilde{y}_{1:T}}^{(i)} \right)^2 \right) \right) \\
& - \frac{1}{2\sigma_a^2} \left(\left(\sigma_a^2{}^{(i)} + \left(\mu_a^{(i)} \right)^2 \right) - 2\mu_a^{(i)} \mu_a + \mu_a^2 \right) - \frac{1}{2} \ln(2\pi\sigma_a^2) \\
& - \ln(\Gamma(a_{\lambda_x})) - a_{\lambda_x} \ln(b_{\lambda_x}) + (a_{\lambda_x} - 1) \left(\psi(a_{\lambda_x}^{(i)}) + \ln(b_{\lambda_x}^{(i)}) \right) - \frac{1}{b_{\lambda_x}} a_{\lambda_x}^{(i)} b_{\lambda_x}^{(i)} \\
& - \frac{1}{2\sigma_b^2} \left(\left(\sigma_b^2{}^{(i)} + \left(\mu_b^{(i)} \right)^2 \right) - 2\mu_b^{(i)} \mu_b + \mu_b^2 \right) - \frac{1}{2} \ln(2\pi\sigma_b^2) \\
& - \ln(\Gamma(a_{\lambda_y})) - a_{\lambda_y} \ln(b_{\lambda_y}) + (a_{\lambda_y} - 1) \left(\psi(a_{\lambda_y}^{(i)}) + \ln(b_{\lambda_y}^{(i)}) \right) - \frac{1}{b_{\lambda_y}} a_{\lambda_y}^{(i)} b_{\lambda_y}^{(i)}
\end{aligned} \tag{6.246}$$

For the entropy term

$$\mathcal{H}(q^{(i)}(\vartheta)) = \mathcal{H}\left(q^{(i)}(a)q^{(i)}(\lambda_x)q^{(i)}(b)q^{(i)}(\lambda_y)q^{(i)}(x_{1:T})\right) \quad (6.247)$$

we first note that we can re-express this term using the additivity of differential entropy for independent variables (Cover and Thomas (2006), p.253) as

$$\mathcal{H}(q^{(i)}(\vartheta)) = \mathcal{H}\left(q^{(i)}(a)\right) + \mathcal{H}\left(q^{(i)}(\lambda_x)\right) + \mathcal{H}\left(q^{(i)}(b)\right) + \mathcal{H}\left(q^{(i)}(\lambda_y)\right) + \mathcal{H}\left(q^{(i)}(x_{1:T})\right) \quad (6.248)$$

The first four terms in this sum correspond to the differential entropies of Gaussian and Gamma probability density functions, whose analytical solutions are readily available from the literature (cf. Supplement Section 2). Specifically, we have

$$\mathcal{H}\left(q^{(i)}(a)\right) = \mathcal{H}\left(N\left(a; \mu_a^{(i)}, \sigma_a^{2(i)}\right)\right) = \frac{1}{2} \ln\left(2\pi e \sigma_a^{2(i)}\right) \quad (6.249)$$

$$\mathcal{H}\left(q^{(i)}(\lambda_x)\right) = \mathcal{H}\left(G\left(\lambda_x; a_{\lambda_x}^{(i)}, b_{\lambda_x}^{(i)}\right)\right) = a_{\lambda_x}^{(i)} + \ln b_{\lambda_x}^{(i)} + \ln \Gamma\left(a_{\lambda_x}^{(i)}\right) + \left(1 - a_{\lambda_x}^{(i)}\right) \psi(a_{\lambda_x}^{(i)}) \quad (6.250)$$

$$\mathcal{H}\left(q^{(i)}(b)\right) = \mathcal{H}\left(N\left(b; \mu_b^{(i)}, \sigma_b^{2(i)}\right)\right) = \frac{1}{2} \ln\left(2\pi e \sigma_b^{2(i)}\right) \quad (6.251)$$

$$\mathcal{H}\left(q^{(i)}(\lambda_y)\right) = \mathcal{H}\left(G\left(\lambda_y; a_{\lambda_y}^{(i)}, b_{\lambda_y}^{(i)}\right)\right) = a_{\lambda_y}^{(i)} + \ln b_{\lambda_y}^{(i)} + \ln \Gamma\left(a_{\lambda_y}^{(i)}\right) + \left(1 - a_{\lambda_y}^{(i)}\right) \psi(a_{\lambda_y}^{(i)}) \quad (6.252)$$

The last term in (2.248) can be evaluated by considering the factorization properties of $q^{(i)}(x_{1:T})$. From (6.3.2.3), we have

$$q^{(i)}(x_{1:T}) = \tilde{p}_{\tilde{\theta}^{(i)}}(x_{1:T}|\tilde{y}_{1:T}) = \tilde{p}_{\tilde{\theta}^{(i)}}(x_1|\tilde{y}_{1:T}) \prod_{t=2}^T \tilde{p}_{\tilde{\theta}^{(i)}}(x_t|x_{t-1}\tilde{y}_{1:T}) = \tilde{p}_{\tilde{\theta}^{(i)}}(x_1|\tilde{y}_{1:T}) \prod_{t=2}^T \frac{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}, x_t|\tilde{y}_{1:T})}{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}|\tilde{y}_{1:T})} \quad (6.253)$$

Substitution in (6.248) thus yields

$$\mathcal{H}\left(q^{(i)}(x_{1:T})\right) = \mathcal{H}\left(\tilde{p}_{\tilde{\theta}^{(i)}}(x_1|\tilde{y}_{1:T}) \prod_{t=2}^T \frac{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}, x_t|\tilde{y}_{1:T})}{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}|\tilde{y}_{1:T})}\right) \quad (6.254)$$

Further, with the additivity of differential entropy for independent variables, we have

$$\begin{aligned} \mathcal{H}\left(q^{(i)}(x_{1:T})\right) &= \mathcal{H}\left(\tilde{p}_{\tilde{\theta}^{(i)}}(x_1|\tilde{y}_{1:T}) \prod_{t=2}^T \frac{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}, x_t|\tilde{y}_{1:T})}{\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}|\tilde{y}_{1:T})}\right) \\ &= \mathcal{H}\left(\tilde{p}_{\tilde{\theta}^{(i)}}(x_1|\tilde{y}_{1:T})\right) + \sum_{t=2}^T \mathcal{H}\left(\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}, x_t|\tilde{y}_{1:T})\right) - \sum_{t=1}^{T-1} \mathcal{H}\left(\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}|\tilde{y}_{1:T})\right) \\ &= \sum_{t=2}^T \mathcal{H}\left(\tilde{p}_{\tilde{\theta}^{(i)}}(x_{t-1}, x_t|\tilde{y}_{1:T})\right) - \sum_{t=2}^{T-1} \mathcal{H}\left(\tilde{p}_{\tilde{\theta}^{(i)}}(x_t|\tilde{y}_{1:T})\right) \end{aligned} \quad (6.255)$$

which, in parameterized form corresponds to

$$\mathcal{H}\left(q^{(i)}(x_{1:T})\right) = \sum_{t=2}^T \mathcal{H}\left(N(x_{t-1:t}; \tilde{\mu}_{x_{t-1:t}|\tilde{y}_{1:T}}, \tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}})\right) - \sum_{t=2}^{T-1} \mathcal{H}\left(N(x_t; \tilde{\mu}_{x_t|\tilde{y}_{1:T}}, \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2)\right)$$

$$\begin{aligned}
&= \frac{1}{2} \sum_{t=2}^T \ln((2\pi e)^2 |\tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}|) - \frac{1}{2} \sum_{t=2}^{T-1} \ln(2\pi e \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2) \\
&= (T-1) \ln(2\pi e) + \sum_{t=2}^T \ln |\tilde{\Sigma}_{x_{t-1:t}|\tilde{y}_{1:T}}| - \frac{T-2}{2} \sum_{t=2}^{T-1} \ln \tilde{\sigma}_{x_t|\tilde{y}_{1:T}}^2
\end{aligned} \tag{6.255}$$

This completes the derivation. $\square$

6.5 Posterior distribution transformation

To estimate a posterior distribution over the latent SDE parameters $\alpha \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$, we used the Euler-Maruyama approximation, which entails the use of two transformation mappings from α and σ onto the LGSSM parameter $a \in \mathbb{R}$ and $\sigma_x^2 \in \mathbb{R}_+$, as discussed in Supplement Section 3. If we obtain posterior pdfs over a and σ_x^2 based on VB inference for the LGSSM and would like to use them for posterior inference over the SDE parameters α and σ , we have to account for the transformation theorem for pdfs, as introduced below.

○ General univariate transformation theorem

For a probability measure P on the Borel σ -field $\mathcal{B}$ with probability density function f for the distribution of the random variable x and a measurable transformation given by

$$T: \mathbb{R} \rightarrow \mathbb{R}, x \mapsto y := T(x) \tag{6.256}$$

a probability density function g for the distribution of the random variable $y = T(x)$ is given by

$$g(y) = \frac{f(T^{-1}(y))}{|T'(T^{-1}(y))|} \tag{6.257}$$

where T' denotes the first derivative and $|\cdot|$ the absolute value function. $\square$

More specifically, for some aspects of the current scenario of interest, we have the

○ Transformation theorem for univariate linear- affine transformation

For the univariate linear-affine transformation case of a probability density function f for the distribution of the random variable x , we have with $a, b \in \mathbb{R}, a \neq 0$

$$T: \mathbb{R} \rightarrow \mathbb{R}, x \mapsto y := T(x) := ax + b \tag{6.258}$$

a probability density function g for the distribution of the random variable $y = T(x)$ given by

$$g(y) = \frac{1}{|a|} \cdot f\left(\frac{y-b}{a}\right) \tag{6.259}$$

$\square$

To transfer probabilistic statements over LGSSM parameters to probabilistic statements over latent SDE system parameters, we will employ the transformation theorems above for univariate transformations below. To this end, we first define the transformations from LGSSM parameters a, λ_x and λ_y to the latent SDE system parameters α, σ and σ_y^2 . Based on the inverse relationships of these mappings established in Supplement Sections 3 and 6, we have

$$T_1: \mathbb{R} \rightarrow \mathbb{R}, a \mapsto T_1(a) = \frac{1}{\Delta t} (a - 1) =: \alpha \tag{6.260}$$

with its inverse (cf. equation (3.53))

$$T_1^{-1}: \mathbb{R} \rightarrow \mathbb{R}, \alpha \mapsto T_1^{-1}(\alpha) = (1 + \alpha\Delta t) =: a \quad (6.261)$$

As we have been working with the reciprocal of σ_x^2 , given by $\lambda_x = \frac{1}{\sigma_x^2}$, we will derive probabilistic statements on the posterior distribution of σ based on

$$T_2: \mathbb{R}_+ \rightarrow \mathbb{R}_+, \lambda_x \mapsto T_2(\lambda_x) = \frac{1}{\sqrt{\lambda_x \Delta t}} =: \sigma \quad (6.262)$$

with inverse (cf. equation (3.53) and the discussion following equation 23 of the tutorial)

$$T_2^{-1}: \mathbb{R}_+ \rightarrow \mathbb{R}_+, \sigma \mapsto T_2^{-1}(\sigma) := \frac{1}{\sigma^2 \Delta t} =: \frac{1}{\sigma_x^2} =: \lambda_x \quad (6.263)$$

Finally, we have for the transformation between λ_y and σ_y^2 (cf. the discussion following equation 23 of the tutorial)

$$T_3: \mathbb{R}_+ \rightarrow \mathbb{R}_+, \lambda_y \mapsto T_3(\lambda_y) = \lambda_y^{-1} =: \sigma_y^2 \quad (6.264)$$

and inverse

$$T_3^{-1}: \mathbb{R}_+ \rightarrow \mathbb{R}_+, \sigma_y^2 \mapsto T_3^{-1}(\sigma_y^2) = (\sigma_y^2)^{-1} = \sigma_y^{-2} =: \lambda_y \quad (6.265)$$

Based on these transformations, the posterior probability density functions f_1, f_2 and f_3 over a, λ_x and λ_y , respectively, correspond to the following posterior probability density functions over α, σ and σ_y^2

$$g_1(\alpha) = \Delta t \cdot f_1(\alpha\Delta t + 1) \quad (6.266)$$

$$g_2(\sigma) = \frac{2f_2\left(\frac{1}{\sigma^2 \Delta t}\right)}{\Delta t \sigma^3} \quad (6.267)$$

and

$$g_3(\sigma_y^2) = (\sigma_y^2)^{-2} f_3\left((\sigma_y^2)^{-1}\right) \quad (6.2xx)$$

○ Proof of equation (6.266)

The transformation of the LGSSM parameter a to the latent SDE system parameter α is given by

$$T_1: \mathbb{R} \rightarrow \mathbb{R}, a \mapsto T_1(a) = \frac{1}{\Delta t}(a - 1) = \frac{1}{\Delta t}a - \frac{1}{\Delta t} \quad (6.268)$$

and the transformation theorem for univariate linear affine transformation applies. Substitution yields

$$g_1(\alpha) = \Delta t \cdot f_1\left(\alpha\Delta t + \frac{\Delta t}{\Delta t}\right) = \Delta t \cdot f_1(\alpha\Delta t + 1) \quad (6.269)$$

□

○ Proof of equation (6.267)

The transformation of the LGSSM parameter λ_x to the latent SDE system parameter σ is given by

$$T_2: \mathbb{R}_+ \rightarrow \mathbb{R}_+, \lambda_x \mapsto T_2(\lambda_x) = (\lambda_x \Delta t)^{-\frac{1}{2}} \quad (6.670)$$

The derivative of T_2 at λ_x evaluates to

$$T_2'(\lambda_x) = \frac{d}{d\lambda_x} \left((\lambda_x \Delta t)^{-\frac{1}{2}} \right) = -\frac{1}{2} \lambda_x^{-\frac{3}{2}} \Delta t^{-\frac{1}{2}} = -\frac{1}{2} \lambda_x^{-\frac{3}{2}} \Delta t^{-\frac{1}{2}} \quad (2.671)$$

Substitution the above and $T_2^{-1}(\sigma) = \frac{1}{\sigma^2 \Delta t}$ in (6.2xx) then yields

$$g_2(\sigma) = \frac{f_2(T_2^{-1}(\sigma))}{|T_2'(T_2^{-1}(\sigma))|} = \frac{f_2\left(\frac{1}{\sigma^2 \Delta t}\right)}{\left|T_2'\left(\frac{1}{\sigma^2 \Delta t}\right)\right|} = \frac{f_2\left(\frac{1}{\sigma^2 \Delta t}\right)}{\left|-\frac{1}{2}\left(\frac{1}{\sigma^2 \Delta t}\right)^{-\frac{3}{2}}(\Delta t)^{-\frac{1}{2}}\right|} = \frac{f_2\left(\frac{1}{\sigma^2 \Delta t}\right)}{\left|-\frac{1}{2}(\sigma^2 \Delta t)^{\frac{3}{2}} \Delta t^{-\frac{1}{2}}\right|} = \frac{f_2\left(\frac{1}{\sigma^2 \Delta t}\right)}{\left|-\frac{1}{2}(\sigma^2)^{\frac{3}{2}}(\Delta t)^{\frac{3}{2}} \Delta t^{-\frac{1}{2}}\right|} = \frac{f_2\left(\frac{1}{\sigma^2 \Delta t}\right)}{\left|-\frac{1}{2}\sigma^3 \Delta t\right|} = \frac{2f_2\left(\frac{1}{\sigma^2 \Delta t}\right)}{\sigma^3 \Delta t} \quad (2.672)$$

□

○ Proof of equation (6.268)

he transformation of the LGSSM parameter λ_y to the latent SDE system parameter σ_y^2 is given by

$$T_3: \mathbb{R}_+ \rightarrow \mathbb{R}_+, \lambda_y \mapsto T_3(\lambda_y) = \lambda_y^{-1} \quad (6.273)$$

The derivative of T_3 at λ_y evaluates to

$$T_3'(\lambda_y) = \frac{d}{d\lambda_y} \lambda_y^{-1} = -\lambda_y^{-2} \quad (6.274)$$

Substitution the above and $T_3^{-1}(\sigma_y^2) = \sigma_y^{-2}$ in (6.2xx) then yields

$$g_3(\sigma_y^2) = \frac{f_3(\sigma_y^{-2})}{|T_3'(\sigma_y^{-2})|} = \frac{f_3(\sigma_y^{-2})}{|-(\sigma_y^{-2})^{-2}|} = \frac{f_3(\sigma_y^{-2})}{(\sigma_y^{-2})^{-2}} = \frac{f_3(\sigma_y^{-2})}{((\sigma_y^2)^{-2})^{-1}} = (\sigma_y^2)^{-2} f_3(\sigma_y^{-2}) = (\sigma_y^2)^{-2} f_3((\sigma_y^2)^{-1}) \quad (6.275)$$

□

6.6 Matlab implementation

The following Matlab code was used to generate Figures 1, 6 and 7 of the tutorial. Note that this code comprises the Kalman filter and KRTS smoothing code depicted previously.

```
function vb_lgssm_2

% This function implements the variational Bayesian estimation of a
% univariate LGSSM/discretized linear latent SDE as discussed in the
% tutorial Section 5 and Supplement Section 6.
%
%
% Copyright (C) Dirk Ostwald
% -----

clc
close all

% -----
% Random number generator settings
% -----
% flag for reproducibility of the JMP figure
isjmpfig = 1;

% set rng
if isjmpfig
    rng(2012873763)
else
    rng('shuffle')
end

% -----
% True, but unknown, model specification and sampling
% -----
% system dimensionality
k = 1 ; % dimensionality of latent variable vector
p = 1 ; % dimensionality of observed variable vector

% SDE parameters
delta_t = 1 ; % time between observations, arbitrary units
T = 50 ; % number of discrete time points
alpha = -0.1 ; % SDE evolution parameter
sigma = 0.1 ; % SDE latent noise parameter

% true, but unknown, parameters
a = delta_t*alpha + 1 ; % LGSSM transition parameter
sigsqr_x = sigma^2*delta_t ; % LGSSM latent variable variance parameter
lambda_x = 1/sigsqr_x ; % LGSSM latent variable precision parameter
b = -200 ; % LGSSM emission parameter
sigsqr_y = 0.01 ; % LGSSM observation noise variance
lambda_y = 1/sigsqr_y ; % LGSSM latent variable precision parameter
c_0 = 100 ; % LGSSM additive constant for reactino time example

% initialize hidden and observed data sample arrays
x = NaN(k,T);
y = NaN(p,T);

% define intial state of the latent variable vector
x(:,1) = -1;

% obtain initial observed state
y_rt(:,1) = mvnrnd((b*(x(:,1)))+ c_0,sigsqr_y);

% cycle over time points t = 2, ..., T
for t = 2:T

    % sample latent state
    x(:,t) = mvnrnd(a*x(:,t-1),sigsqr_x);

    % sample observed state
    y_rt(:,t) = mvnrnd((b*(x(:,t)))+ c_0,sigsqr_y);

end
y = y_rt - c_0;
```

```

% -----
%                                     Generative model formulation
% -----
% prior distribution parameters
mu_1      = -1      ;
lambda_1  = 1000    ;
sigsqr_1  = 1/lambda_x ;
m_a       = 10      ;
s_a       = 1e3     ;
a_lx      = 1e-1    ;
b_lx      = 1e3     ;
m_b       = -190    ;
s_b       = 1e3     ;
a_ly      = 1e0     ;
b_ly      = 1e2     ;

% concatenation of priors for variational free energy evaluation
prior.mu_1 = mu_1      ;
prior.lambda_1 = lambda_1 ;
prior.m_a     = m_a     ;
prior.s_a     = s_a     ;
prior.a_lx    = a_lx    ;
prior.b_lx    = b_lx    ;
prior.m_b     = m_b     ;
prior.s_b     = s_b     ;
prior.a_ly    = a_ly    ;
prior.b_ly    = b_ly    ;

% -----
%                                     Parameter Space Initialization
% -----
% specify linearly partitioned parameter spaces for inference
amin      = -300;
amax      = 300;
ares      = 10e4;
aspace    = linspace(amin, amax, ares);

lxmax     = 200;
lxmin     = 0;
lxres     = 10e4;
lxspace   = linspace(lxmin, lxmax, lxres);

bmin      = -300;
bmax      = 300;
bres      = 10e4;
bspace    = linspace(bmin, bmax, bres);

lymax     = 200;
lymin     = 0;
lyres     = 10e4;
lyspace   = linspace(lymin, lymax, lyres);

% -----
%                                     Generative model estimation/inversion
% -----
% maximum number of iterations
max_i     = 5;

% initialize variational parameter arrays
q_m_a     = NaN(max_i,1);
q_s_a     = NaN(max_i,1);
q_a_lx    = NaN(max_i,1);
q_b_lx    = NaN(max_i,1);
q_m_b     = NaN(max_i,1);
q_s_b     = NaN(max_i,1);
q_a_ly    = NaN(max_i,1);
q_b_ly    = NaN(max_i,1);
q_mu_x    = NaN(max_i,T);
q_sig_x   = NaN(max_i,T);
q_sig_xx  = NaN(max_i,T-1);

% initialize variational free energy array
FE        = NaN(max_i,1);

% initialize average energy and entropy component arrays
n_ae      = 14      ; % number of average energy terms
n_h       = 7       ; % number of entropy terms
E         = NaN(n_ae,max_i) ;
H         = NaN(n_h,max_i)  ;

```

```

% VB algorithm initialization
% -----
% variational parameter initialization for a,b
q_m_a(1)      = m_a;
q_s_a(1)      = s_a;
q_a_lx(1)     = a_lx;
q_b_lx(1)     = b_lx;
q_m_b(1)      = m_b;
q_s_b(1)      = s_b;
q_a_ly(1)     = a_ly;
q_b_ly(1)     = b_ly;

% variational parameter initialization for x_(1:T)
y_til        = ly ; ...
              zeros(k,T) ; ...
              zeros(p,T) ] ;
a_til        = q_m_a(1) ;
sigsqr_x_til = (q_a_lx(1)*q_b_lx(1))^( -1) ;
b_til_t      = [q_m_b(1) ; ...
               sqrt(q_s_a(1)*q_a_lx(1)*q_b_lx(1)) ; ...
               sqrt(q_s_b(1)*q_a_ly(1)*q_b_ly(1)) ] ;
b_til_T      = [q_m_b(1) ; ...
               0 ; ...
               sqrt(q_s_b(1)*q_a_ly(1)*q_b_ly(1))] ;
sigsqr_y_til = [(q_a_ly(1)*q_b_ly(1))^( -1) 0 0 ; ...
               0 1 0 ; ...
               0 0 1 ] ;

[mu_f_til, sigma_f_til] = kalman_filter(y_til, mu_1, sigsqr_1, a_til, sigsqr_x_til, b_til_t, b_til_T,
sigsqr_y_til);
[mu_s_til, sigma_s_til, sigma_cs_til] = krts_smoother(y, mu_f_til, a_til, sigma_f_til, sigsqr_x_til);

q_mu_x(1,:) = squeeze(mu_s_til) ;
q_sig_x(1,:) = squeeze(sigma_s_til) ;
q_sig_xx(1,:) = squeeze(sigma_cs_til) ;

% concatenate variational parameters for variational free energy evaluation
varia.q_m_a = q_m_a(1) ;
varia.q_s_a = q_s_a(1) ;
varia.q_a_lx = q_a_lx(1) ;
varia.q_b_lx = q_b_lx(1) ;
varia.q_m_b = q_m_b(1) ;
varia.q_s_b = q_s_b(1) ;
varia.q_a_ly = q_a_ly(1) ;
varia.q_b_ly = q_b_ly(1) ;
varia.q_mu_x = q_mu_x(1,:) ;
varia.q_sig_x = q_sig_x(1,:) ;
varia.q_sig_xx = q_sig_xx(1,:) ;

% evaluate variational free energy
vFE = vFreeEnergy(y,varia,prior);
FE(1) = vFE.F;
E(:,1) = vFE.E';
H(:,1) = vFE.H';

% VB algorithm iterations
% -----
for i = 1:max_i-1

    % update of q(a)
    % -----
    % \sum_{t=1}^{T-1} <x_t^2>_q(x_t)
    txtxt_qxt = sum(q_sig_x(i,1:T-1),2) + q_mu_x(i,1:T-1)*q_mu_x(i,1:T-1)';

    % \sum_{t=1}^{T-1} <x_t, x_{t+1}>_q(x_t:t+1)
    txttt_qxt = sum(q_sig_xx(i,1:T-1),2) + q_mu_x(i,1:T-1)*q_mu_x(i,2:T)';

    % update equations
    q_s_a(i+1) = (lambda_x*(txtxt_qxt + (1/s_a)))^( -1);
    q_m_a(i+1) = q_s_a(i+1)*((lambda_x*txttt_qxt) + (m_a/s_a));

    % update of q(\lambda_x)
    % -----
    % \sum_{t=2}^T <x_t^2>_q(x_t)
    txtxt_qxt_1 = sum(q_sig_x(i,2:T),2) + q_mu_x(i,2:T)*q_mu_x(i,2:T)';

    % \sum_{t=1}^{T-1} <x_t, x_{t+1}>_q(x_t:t+1)
    txttt_qxt = sum(q_sig_xx(i,1:T-1),2) + q_mu_x(i,1:T-1)*q_mu_x(i,2:T)';

    % \sum_{t=1}^{T-1} <x_t^2>_q(x_t)
    txtxt_qxt_2 = sum(q_sig_x(i,1:T-1),2) + q_mu_x(i,1:T-1)*q_mu_x(i,1:T-1)';

```

```

% update equations
q_a_lx(i+1) = (T/2) + a_lx;
q_b_lx(i+1) = ((1/b_lx) + 1/2*(xtxt_qxt_1 - 2*q_m_a(i+1)*xtxt_qxt + (q_s_a(i+1) +
q_m_a(i+1)^2)*xtxt_qxt_2))^( -1);

% update of q(b)
% -----
% \sum_{t=1}^T <x_t^2>_q(x_t)
xtxt_qxt = sum(q_sig_x(i,1:T),2) + q_mu_x(i,1:T)*q_mu_x(i,1:T)';

% \sum_{t=1}^T y_t<x_t>_q(x_t)
ytxt_qxt = y*q_mu_x(i,:)';

% update equations
q_s_b(i+1) = (q_a_ly(i)*q_b_ly(i)*xtxt_qxt + (1/s_b))^( -1) ;
q_m_b(i+1) = q_s_b(i+1)*(q_a_ly(i)*q_b_ly(i)*ytxt_qxt + (m_b/s_b));

% update of q(\lambda_y)
% -----
% \sum_{t=1}^T y_t^2
ytyt = y*y';

% \sum_{t=1}^T y_t<x_t>_q(x_t)
ytxt_qxt = y*q_mu_x(i,:)';

% update equations
q_a_ly(i+1) = (T/2) + a_ly;
q_b_ly(i+1) = ((1/b_ly) + 1/2*(ytyt - 2*q_m_b(i+1)*ytxt_qxt + (q_s_b(i+1) + q_m_b(i+1)^2)*xtxt_qxt))^( -
1);

% update of q(x_{1:T})
% -----
y_til = [y ; ...
zeros(k,T) ; ...
zeros(p,T) ] ;
a_til = q_m_a(i+1) ;
sigsqr_x_til = (q_a_lx(i+1)*q_b_lx(i+1))^( -1) ;
b_til_t = [q_m_b(i+1) ; ...
sqrt(q_s_a(i+1)*q_a_lx(i+1)*q_b_lx(i+1)) ; ...
sqrt(q_s_b(i+1)*q_a_ly(i+1)*q_b_ly(i+1)) ] ;
b_til_T = [q_m_b(i+1) ; ...
0 ; ...
sqrt(q_s_b(i+1)*q_a_ly(i+1)*q_b_ly(i+1)) ] ;
sigsqr_y_til = [(q_a_ly(i+1)*q_b_ly(i+1))^( -1) 0 0 ; ...
0 1 0 ; ...
0 0 1 ] ;

% q(x_{1:T}) parameter inference
[mu_f_til, sigma_f_til] = kalman_filter(y_til, mu_1, sigsqr_1, a_til, sigsqr_x_til,
b_til_t, b_til_T, sigsqr_y_til);
[mu_s_til, sigma_s_til, sigma_cs_til] = krts_smoother(y, mu_f_til, a_til, sigma_f_til, sigsqr_x_til);
q_mu_x(i+1,:) = squeeze(mu_s_til) ;
q_sig_x(i+1,:) = squeeze(sigma_s_til) ;
q_sig_xx(i+1,:) = squeeze(sigma_cs_til) ;

% concatenate variational parameters for variational free energy evaluation
varia.q_m_a = q_m_a(i+1) ;
varia.q_s_a = q_s_a(i+1) ;
varia.q_a_lx = q_a_lx(i+1) ;
varia.q_b_lx = q_b_lx(i+1) ;
varia.q_m_b = q_m_b(i+1) ;
varia.q_s_b = q_s_b(i+1) ;
varia.q_a_ly = q_a_ly(i+1) ;
varia.q_b_ly = q_b_ly(i+1) ;
varia.q_mu_x = q_mu_x(i+1,:) ;
varia.q_sig_x = q_sig_x(i+1,:) ;
varia.q_sig_xx = q_sig_xx(i+1,:) ;

% evaluate variational free energy
vFE = vFreeEnergy(y,varia,prior);
FE(i+1) = vFE.F;
E(:,i+1) = vFE.E';
H(:,i+1) = vFE.H';

```

end

```

% plot perceptual learning example I
% -----
h = figure;
set(h, 'Color', [1 1 1])
plot([1:T], y_rt, 'LineStyle', '--', 'Color', 'k', 'LineWidth', 3, 'Marker', 'o', 'MarkerFaceColor', 'k',
'MarkerSize', 12);
ylabel('Reaction Time [ms]', 'FontName', 'Times New Roman', 'FontSize', 50)
xlabel('Time [trials]', 'FontName', 'Times New Roman', 'FontSize', 50)
set(gca, 'FontSize', 50, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [1 T], 'YLim', [0
300], 'YColor', 'k', 'YTick', [100 200 300])
grid on
maximize(h)

% plot perceptual learning example II
% -----
h = figure;
set(h, 'Color', [1 1 1])
hold on
[ax,h1,h2] = plotyy([1:T], y_rt, [1:T], q_mu_x(3,:), 'plot');
set(get(ax(1), 'Ylabel'), 'String', 'Reaction Time [ms]', 'FontName', 'Times New Roman', 'FontSize', 50)
set(get(ax(2), 'Ylabel'), 'String', 'Cognitive State[a.u.]', 'FontName', 'Times New Roman', 'FontSize', 50)
set(h1, 'LineStyle', '--', 'Color', 'k', 'LineWidth', 3, 'Marker', 'o', 'MarkerFaceColor', 'k', 'MarkerSize',
12)
set(h2, 'LineStyle', '--', 'Color', 'r', 'LineWidth', 3, 'Marker', 'o', 'MarkerFaceColor', 'r', 'MarkerSize',
12)
xlabel('Time [trials]', 'FontName', 'Times New Roman', 'FontSize', 50)
set(ax(1), 'FontSize', 50, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [1 T], 'YLim', [0
310], 'YColor', 'k', 'YTick', [100 200 300])
set(ax(2), 'FontSize', 50, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [1 T], 'YLim', [-2 2 ],
'YColor', 'r', 'YTick', -2:2)

[ax,h3,h4] = plotyy([1:T], y_rt, [1:T], x
, 'plot');
set(get(ax(1), 'Ylabel'), 'String', 'Reaction Time [ms]', 'FontName', 'Times New Roman', 'FontSize', 50)
set(get(ax(2), 'Ylabel'), 'String', 'Cognitive State[a.u.]', 'FontName', 'Times New Roman', 'FontSize', 50)
set(h3, 'LineStyle', '--', 'Color', 'k', 'LineWidth', 3, 'Marker', 'o', 'MarkerFaceColor', 'k', 'MarkerSize',
12)
set(h4, 'LineStyle', '--', 'Color', 'k', 'LineWidth', 1, 'Marker', 'o', 'MarkerFaceColor', 'w', 'MarkerSize',
10)
xlabel('Time [trials]', 'FontName', 'Times New Roman', 'FontSize', 50)
set(ax(1), 'FontSize', 50, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [1 T], 'YLim', [0
310], 'YColor', 'k', 'YTick', [100 200 300])
set(ax(2), 'FontSize', 50, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [1 T], 'YLim', [-2 2 ],
'YColor', 'r', 'YTick', -2:2)
grid on
maximize(h)

% plot variational distribution updates
% -----
h = figure;
set(h, 'Color', [1 1 1])
iter = [1 2 5];

% cycle over the first three iterations
for i = 1:3

    % q(a) updates
    subplot(6,3,i)
    hold on
    plot(aspace, pdf('Norm', aspace, q_m_a(iter(i)), sqrt(q_s_a(iter(i)))), 'b', 'LineWidth', 2)
    plot(a, 0, 'ro', 'MarkerFaceColor', 'r', 'Markersize', 10)
    plot(q_m_a(iter(i)), 0, 'ko', 'MarkerFaceColor', 'w', 'Markersize', 10)
    set(gca, 'LineWidth', 1, 'FontName', 'Times New Roman', 'FontSize', 16)
    if i == 1
        xlim([0 20])
        ylim([0 0.05])
    else
        xlim([0 2])
        ylim([0 14])
    end

    % q(\lambda_x) updates
    subplot(6,3,i+3)
    hold on
    plot(lxspace, pdf('Gam', lxspace, q_a_lx(iter(i)), q_b_lx(iter(i))), 'b', 'LineWidth', 2)
    plot(lambda_x, 0, 'ro', 'MarkerFaceColor', 'r', 'Markersize', 10)
    plot(q_a_lx(iter(i))*q_b_lx(iter(i)), 0, 'ko', 'MarkerFaceColor', 'w', 'Markersize', 10)
    set(gca, 'LineWidth', 1, 'FontName', 'Times New Roman', 'FontSize', 16)
    ylim([0 0.05])

```

```

% q(b) updates
subplot(6,3,i+6)
hold on
plot(bspace, pdf('Norm', bspace, q_m_b(iter(i)), sqrt(q_s_b(iter(i)))), 'b', 'LineWidth', 2)
plot(b, 0, 'ro', 'MarkerFaceColor', 'r', 'Markersize', 10)
plot(q_m_b(iter(i)), 0, 'ko', 'MarkerFaceColor', 'w', 'Markersize', 10)
set(gca, 'LineWidth', 1, 'FontName', 'Times New Roman', 'FontSize', 16)
if i == 1
    xlim([-220 -170])
    ylim([0 0.05])
else
    xlim([-202 -198])
    ylim([0 12])
end

% q(\lambda_y) updates
subplot(6,3,i+9)
hold on
plot(lyspace, pdf('Gam', lyospace, q_a_ly(iter(i)), q_b_ly(iter(i))), 'b', 'LineWidth', 2)
plot(lambda_y, 0, 'ro', 'MarkerFaceColor', 'r', 'Markersize', 10)
plot(q_a_ly(iter(i))*q_b_ly(iter(i)), 0, 'ko', 'MarkerFaceColor', 'w', 'Markersize', 10)
set(gca, 'LineWidth', 1, 'FontName', 'Times New Roman', 'FontSize', 16)

% q(x_1:T) update
subplot(6,3,i+12)
hold on
plot([1:T], x, 'ro-', 'MarkerSize', 5, 'MarkerFaceColor', 'r', 'LineWidth', 1)
plot([1:T], q_mu_x(iter(i),:), 'ko-', 'MarkerSize', 5, 'MarkerFaceColor', 'w', 'LineWidth', 1)
xlim([0 T])
ylim([-1 1])
set(gca, 'LineWidth', 1, 'FontName', 'Times New Roman', 'FontSize', 16)

end

% variational free energy
subplot(6,3,[16 18])
plot([0:max_i-1], FE, 'ko-', 'MarkerFaceColor', 'w', 'LineWidth', 2, 'MarkerSize', 8)
box off
set(gca, 'LineWidth', 1, 'FontName', 'Times New Roman', 'FontSize', 16, 'xtick', [0:max_i-1])
xlim([-1 max_i])
maximize(h)

% plot SDE posterior parameter distributions
% -----
% transform posterior pdf over a to posterior pdf over \alpha
alphaspace = linspace(-2,2,5000);
p_alpha_pri = NaN(length(alphaspace),1);
p_alpha_pos = NaN(length(alphaspace),1);
for i = 1:length(alphaspace)
    p_alpha_pri(i) = delta_t*pdf('Norm', (delta_t*alphaspace(i)) + 1, m_a, sqrt(s_a));
    p_alpha_pos(i) = delta_t*pdf('Norm', (delta_t*alphaspace(i)) + 1, q_m_a(end), sqrt(q_s_a(end)));
end

% transform posterior pdf over \lambda_x to posterior pdf over sigma
sigmaspace = linspace(eps,.3,5000);
p_sigma_pri = NaN(length(sigmaspace),1);
p_sigma_pos = NaN(length(sigmaspace),1);

for i = 1:length(sigmaspace)
    p_sigma_pri(i) = (2/(delta_t*(sigmaspace(i)^3)))*pdf('Gam', 1/(delta_t*(sigmaspace(i)^2)), a_lx, b_lx);
    p_sigma_pos(i) = (2/(delta_t*(sigmaspace(i)^3)))*pdf('Gam', 1/(delta_t*(sigmaspace(i)^2)), q_a_lx(end), q_b_lx(end));
end

% no transformation necessary for b
p_b_pri = pdf('Norm', bspace, m_b, sqrt(s_b));
p_b_pos = pdf('Norm', bspace, q_m_b(end), sqrt(q_s_b(end)));

% transform posterior pdf over \lambda_y to posterior pdf over sigma_y
sigsqryspace = linspace(eps,0.2,1000);
p_sigsqr_y_pri = NaN(length(sigsqryspace),1);
p_sigsqr_y_pos = NaN(length(sigsqryspace),1);

for i = 1:length(sigsqryspace)
    p_sigsqr_y_pri(i) = (sigsqryspace(i)^(-2))*pdf('Gam', (sigsqryspace(i)^(-1)), a_ly, b_ly);
    p_sigsqr_y_pos(i) = (sigsqryspace(i)^(-2))*pdf('Gam', (sigsqryspace(i)^(-1)), q_a_ly(end), q_b_ly(end));
end

```

```

% plot the posterior distributions over SDE parameter a
h = figure;
set(h, 'Color', [1 1 1])
subplot(2,2,1)
[ax,h1,h2] = plotyy(alphaspace ,p_alpha_pri, alphaspace , p_alpha_pos, 'plot');
set(get(ax(1),'Ylabel') , 'String','Prior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(get(ax(2),'Ylabel') , 'String','Posterior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(h1,'LineStyle','-','LineWidth', 3)
set(h2,'LineStyle','-','LineWidth', 3)
xlabel('\alpha', 'FontName', 'Times New Roman', 'FontSize', 40)
set(ax(1), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [-1 1])
set(ax(2), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [-1 1])

% plot the posterior distributions over SDE parameter sigma
subplot(2,2,2)
[ax,h1,h2] = plotyy(sigmaspace ,p_sigma_pri, sigmaspace , p_sigma_pos, 'plot');
set(get(ax(1),'Ylabel') , 'String','Prior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(get(ax(2),'Ylabel') , 'String','Posterior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(h1,'LineStyle','-','LineWidth', 3)
set(h2,'LineStyle','-','LineWidth', 3)
xlabel('\sigma', 'FontName', 'Times New Roman', 'FontSize', 40)
set(ax(1), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [sigmaspace(1)
sigmaspace(end)])
set(ax(2), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [sigmaspace(1)
sigmaspace(end)])

% plot the posterior distributions over parameter b
subplot(2,2,3)
[ax,h1,h2] = plotyy(bspace, p_b_pri, bspace , p_b_pos, 'plot');
set(get(ax(1),'Ylabel') , 'String','Prior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(get(ax(2),'Ylabel') , 'String','Posterior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(h1,'LineStyle','-','LineWidth', 3)
set(h2,'LineStyle','-','LineWidth', 3)
xlabel('b', 'FontName', 'Times New Roman', 'FontSize', 40)
set(ax(1), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [-210 -190])
set(ax(2), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [-210 -190])

% plot the posterior distributions over parameter \sigma_y^2
subplot(2,2,4)
[ax,h1,h2] = plotyy(sigsqrspace , p_sigsqr_y_pri, sigsqrspace , p_sigsqr_y_pos, 'plot');
set(get(ax(1),'Ylabel') , 'String','Prior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(get(ax(2),'Ylabel') , 'String','Posterior' , 'FontName', 'Times New Roman', 'FontSize', 26)
set(h1,'LineStyle','-','LineWidth', 3)
set(h2,'LineStyle','-','LineWidth', 3)
xlabel('\sigma_y^2', 'FontName', 'Times New Roman', 'FontSize', 40)
set(ax(1), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [sigsqrspace(1)
sigsqrspace(end)])
set(ax(2), 'FontSize', 26, 'FontName', 'Times New Roman', 'LineWidth', 3, 'XLim', [sigsqrspace(1)
sigsqrspace(end)])
maximize(h)

% display alpha credible interval
% -----
% numeric integration scheme for the evaluation of a credible interval
delta_alpha = diff(alphaspace) ;
cdf_alpha = NaN(1,length(delta_alpha)) ;
cdf_alpha(1) = 0 ;
for i = 2:length(cdf_alpha)
    cdf_alpha(i) = cdf_alpha(i-1) + (delta_alpha(i-1)*((p_alpha_pos(i) + p_alpha_pos(i-1))/2));
end

% find first index with p(\alpha) > 0.025
lower_idx = 1;
while cdf_alpha(lower_idx) < 0.025
    lower_idx = lower_idx + 1;
end

% find last index with p(\alpha) < 0.975
upper_idx = 1;
while cdf_alpha(upper_idx) < 0.975
    upper_idx = upper_idx + 1;
end
upper_idx = upper_idx - 1;

disp([num2str(cdf_alpha(upper_idx) - cdf_alpha(lower_idx),2) ' credible interval: [' , ...
    num2str(alphaspace(lower_idx),2), ' , ' num2str(alphaspace(upper_idx),2), ' ]'])

% display sigma credible interval
% -----
% numeric integration scheme for the evaluation of a credible interval
delta_sigma = diff(sigmaspace) ;

```

```

cdf_sigma      = NaN(1,length(delta_sigma))    ;
cdf_sigma(1) = 0                               ;
for i = 2:length(cdf_sigma)
    cdf_sigma(i) = cdf_sigma(i-1) + (delta_sigma(i-1)*((p_sigma_pos(i) + p_sigma_pos(i-1))/2));
end

% find first index with p(\sigma) > 0.025
lower_idx = 1;
while cdf_sigma(lower_idx) < 0.025
    lower_idx = lower_idx + 1;
end

% find last index with p(\sigma) < 0.975
upper_idx = 1;
while cdf_sigma(upper_idx) < 0.975
    upper_idx = upper_idx + 1;
end
upper_idx = upper_idx - 1;

disp([num2str(cdf_sigma(upper_idx) - cdf_sigma(lower_idx),2) ' credible interval: [' , ...
    num2str(sigmaspace(lower_idx),2), ' , ' num2str(sigmaspace(upper_idx),2), ']]')

end

function [mu_f,Sigma_f] = kalman_filter(y,mu_x_1,Sigma_x_1,A,Sigma_x,B_t,B_T,Sigma_y)

% This function implements the forward pass for inference in a linear
% Gaussian state space model, also known as the Kalman Filter. It
% capitalizes on the expressions for the conditional expectation and
% covariance as discussed in section 5 of the tutorial. Specifically, for
% the covariance update, Joseph's symmetrized form is employed.
%
% Inputs
%
%     y           : p x T data array
%     mu_x_1      : k x 1 initial latent variable expectation
%     Sigma_x_1   : k x k initial latent variable covariance
%     A           : k x k transition matrix
%     Sigma_x     : k x k latent variable noise covariance
%     B_t         : p x k emission matrix for t = 1,...,T-1
%     B_T         : p x k emission matrix for t = T
%     Sigma_y     : p x p latent variable noise covariance
%
% Outputs
%
%     mu_f        : k x T      filtered mean of p(x_t|y_(1:t))
%     Sigma_f     : k x k x T  filtered covariance of p(x_t|y_(1:t))
%
% Copyright (C) Dirk Ostwald
% -----
% recover system and data dimensionalities
k      = size(A,1);
p      = size(y,1);
T      = size(y,2);

% initialize conditional expectation and covariance arrays
mu_f   = NaN(k,T);
Sigma_f = NaN(k,k,T);

% Initialization
% -----
B      = B_t;
mu_f(:,1) = mu_x_1 + Sigma_x_1*B'*inv(B*Sigma_x_1*B' + Sigma_y)*(y(:,1) - B*mu_x_1);
Sigma_f(:, :, 1) = Sigma_x_1 - Sigma_x_1*B'*inv(B*Sigma_x_1*B' + Sigma_y)*B*Sigma_x_1;

% filtered covariance symmetry check
if ~isequal(Sigma_f(:, :, 1),Sigma_f(:, :, 1)')
    Sigma_f(:, :, 1) = 1/2*(Sigma_f(:, :, 1) + Sigma_f(:, :, 1)');
end

% Recursion
% -----
% cycle over remaining time-points
for t = 2:T

    % case distinction for augmented systems
    if t == T
        B = B_T;
    else
        B = B_t;
    end
end

```

```

% predictor and Kalman gain matrices
P = A*Sigma_f(:, :, t-1)*A' + Sigma_x;
K = P*B'*inv(B*P*B' + Sigma_y);

% filtered expectation and covariance updates, Joseph's form
mu_f(:, t) = A*mu_f(:, t-1) + K*(y(:, t) - B*A*mu_f(:, t-1));
Sigma_f(:, :, t) = (eye(k) - K*B)*P*(eye(k) - K*B)' + K*Sigma_y*K';

% filtered covariance symmetry check
if ~isequal(Sigma_f(:, :, t), Sigma_f(:, :, t)')
    Sigma_f(:, :, t) = 1/2*(Sigma_f(:, :, t) + Sigma_f(:, :, t)');
end
end
end

function [mu_s, Sigma_s, Sigma_cs] = krts_smoother(y, mu_f, A, Sigma_f, Sigma_x)

% This function implements the backward pass for inference in a linear
% Gaussian state space model, also known as the Kalman-Rauch-Tung-Striebel
% smoother, and here in addition returning the cross-moment covariance of
%  $p(x_t, x_{t+1} | y_{1:T})$ 
%
% Inputs
%
% y : p x T data array
% mu_f : k x T filtered latent variable expectation
% A : k x k transition matrix
% Sigma_f : k x k filtered latent variable covariance
% Sigma_x : k x k latent variable noise covariance
%
% Outputs
%
% mu_s : k x 1 smoothed mean of  $p(x_t | y_{1:T})$ 
% Sigma_s : k x k x T filtered covariance of  $p(x_t | y_{1:T})$ 
% Sigma_cs : k x k x T-1 filtered cross moment covariance of
%            $p(x_t, x_{t+1} | y_{1:T})$  for  $t = 1, \dots, T$ 
%
% Copyright (C) Dirk Ostwald
% -----
% recover system and data dimensionalities
k = size(A, 1);
p = size(y, 1);
T = size(y, 2);

% initialize conditional expectation, covariance and cross moment
% covariance arrays
mu_s = NaN(k, T);
Sigma_s = NaN(k, k, T);
Sigma_cs = NaN(k, k, T-1);

% filter initialization
% -----
mu_s(:, T) = mu_f(:, T);
Sigma_s(:, :, T) = Sigma_f(:, :, T);

% filter recursion
% -----
for t = T-1:-1:1

    % smoothing matrix J
    J = Sigma_f(:, :, t)*A'*inv(A*Sigma_f(:, :, t)*A' + Sigma_x);

    % smoothed expectation and covariance updates
    mu_s(:, t) = mu_f(:, t) + J*(mu_s(:, t+1) - A*mu_f(:, t));
    Sigma_s(:, :, t) = Sigma_f(:, :, t) + J*(Sigma_s(:, :, t+1) - Sigma_f(:, :, t+1))*J';

    % smoothed covariance symmetry check
    if ~isequal(Sigma_s(:, :, t), Sigma_s(:, :, t)')
        Sigma_s(:, :, t) = 1/2*(Sigma_s(:, :, t) + Sigma_s(:, :, t)');
    end

    % cross moment
    Sigma_cs(:, :, t) = J*Sigma_s(:, :, t+1);

end
end

```

```

function [vFE] = vFreeEnergy(y,varia,prior)

% This function evaluates the variational free energy of the tutorial
% example generative model and its variational approximation
%
% Inputs
%       y       : 1 x T array of observed variable values
%       varia    : structure comprising the variational parameters
%       prior    : structure comprising the prior parameters
%
% Outputs
%       vFE      : variational free energy
%
% Copyright (C) Dirk Ostwald
% -----
% extract variational and prior parameters
q_m_a      = varia.q_m_a ;
q_s_a      = varia.q_s_a ;
q_a_lx     = varia.q_a_lx ;
q_b_lx     = varia.q_b_lx ;
q_m_b      = varia.q_m_b ;
q_s_b      = varia.q_s_b ;
q_a_ly     = varia.q_a_ly ;
q_b_ly     = varia.q_b_ly ;
q_mu_x     = varia.q_mu_x ;
q_sig_x    = varia.q_sig_x ;
q_sig_xx   = varia.q_sig_xx;

m_a        = prior.m_a ;
s_a        = prior.s_a ;
a_lx       = prior.a_lx ;
b_lx       = prior.b_lx ;
m_b        = prior.m_b ;
s_b        = prior.s_b ;
a_ly       = prior.a_ly ;
b_ly       = prior.b_ly ;
mu_l       = prior.mu_l ;
lambda_1   = prior.lambda_1;

% evaluate further constants
T          = size(y,2)
xTxT_qxT   = sum(q_sig_x(2:T),2) + q_mu_x(2:T)*q_mu_x(2:T)' ; % \sum_(t=2)^(T) <x_t^2>_q(x_t)
xtxtt_qxt  = sum(q_sig_xx(1:T-1),2) + q_mu_x(1:T-1)*q_mu_x(2:T)' ; % \sum_(t=1)^(T-1)
<x_t,x_{t+1}>_q(x_{t+1})
xtxt_qxt   = sum(q_sig_x(1:T-1),2) + q_mu_x(1:T-1)*q_mu_x(1:T-1)' ; % \sum_(t=1)^(T-1) <x_t^2>_q(x_t)
ytyt       = y*y' ; % \sum_(t=1)^(T) y_t^2
ytxt_qxt   = y(2:T)*q_mu_x(2:T)' ; % \sum_(t=2)^(T) y_t*<x_t>_q(x_t)
xTTxTT_qxTT = sum(q_sig_x(1:T),2) + q_mu_x(1:T)*q_mu_x(1:T)' ; % \sum_(t=1)^(T) <x_t^2>_q(x_t)

% initialize (x_{t-1},x_t) distribution covariance matrix
log_det_Sigma_xtxtt = NaN(1,T-1);
for t = 2:T

    % assemble covariance matrices and evaluate their determinants
    log_det_Sigma_xtxtt(t-1) = log(det([q_sig_x(t-1) q_sig_xx(t-1);q_sig_xx(t-1) q_sig_x(t)]));

end

% evaluate free energy "average energy" terms
vFE.E(1) = +0.5*log(lambda_1/(2*pi));
vFE.E(2) = -0.5*lambda_1*((q_sig_x(1) + q_mu_x(1)^2) - (2*q_mu_x(1)*mu_l) + (mu_l^2));
vFE.E(3) = +0.5*T*(psi(q_a_lx) + log(q_b_lx)) - 0.5*T*log(2*pi) ;
vFE.E(4) = -0.5*(q_a_lx*q_b_lx)*(xTxT_qxT - (2*q_m_a*xtxtt_qxt) + ((q_s_a + q_m_a^2)*xtxt_qxt));
vFE.E(5) = +0.5*T*(psi(q_a_ly) + log(q_b_ly)) - 0.5*T*log(2*pi) ;
vFE.E(6) = -0.5*(q_a_ly*q_b_ly)*(ytyt - (2*q_m_b*ytxt_qxt) + ((q_s_b + q_m_b^2)*xTTxTT_qxTT));
vFE.E(7) = -(0.5/s_a)*((q_s_a + q_m_a^2) - 2*q_m_a*m_a + m_a^2);
vFE.E(8) = -0.5*log(2*pi*s_a);
vFE.E(9) = -gammaln(a_lx) - (a_lx*log(b_lx));
vFE.E(10) = +(a_lx - 1)*(psi(q_a_lx) + log(q_b_lx)) - ((q_a_lx*q_b_lx)/b_lx);
vFE.E(11) = -(0.5/s_b)*((q_s_b + q_m_b^2) - 2*q_m_b*m_b + m_b^2);
vFE.E(12) = -0.5*log(2*pi*s_b);
vFE.E(13) = -gammaln(a_ly) - (a_ly*log(b_ly));
vFE.E(14) = +(a_ly - 1)*(psi(q_a_ly) + log(q_b_ly)) - ((q_a_ly*q_b_ly)/b_ly);

```

```

% evaluate the free energy "entropy" terms
vFE.H(1) = 0.5*log(2*pi*exp(1)*q_s_a);
vFE.H(2) = q_a_lx + log(q_b_lx) + gammaln(q_a_lx) + (1 - q_a_lx)*psi(q_a_lx);
vFE.H(3) = 0.5*log(2*pi*exp(1)*q_s_b);
vFE.H(4) = q_a_ly + log(q_b_ly) + gammaln(q_a_ly) + (1 - q_a_ly)*psi(q_a_ly);
vFE.H(5) = (T-1)*log(2*pi*exp(1));
vFE.H(6) = sum(log_det_Sigma_txttt);
vFE.H(7) = -((T-1)/2)*sum(log(q_sig_x(2:T-1)));

% evaluate free energy
vFE.F = sum(vFE.E) + sum(vFE.H);

end

function maximize(fig)

% This function maximizes the figure with handle fig
%
% Input
%      fig      : figure handle of figure to be maximized
%
% Original author: Bill Finger, Creare Inc.
% Free for redistribution as long as credit comments are preserved.
%
% Copyright (C) Dirk Ostwald
% -----
if nargin == 0
    fig = gcf;
end

units = get(fig,'units');

set(fig,'units','normalized','outerposition',[0 0 1 1]);
set(fig,'units',units);

end

```